

A Quantitative and High-Throughput Approach to Gene Regulation in *Escherichia coli*

Thesis by
William Thornton Ireland

In Partial Fulfillment of the Requirements for the
Degree of
Doctorate of Philosophy in Physics



CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2020
Defended March 3, 2020

© 2020

William Thornton Ireland
ORCID: 0000-0003-0971-2904

Some rights reserved. This thesis is distributed under a Creative Commons
Attribution License.

ACKNOWLEDGEMENTS

Firstly I'd like to thank my advisor Rob Phillips for his always wonderful scientific insights and attitude. You have taught me so much. I'd also like to thank my lab mates Suzy Beeler, Nathan Belliveau, and Stephanie Barnes, for all the interesting conversations, ideas, and help throughout the years. Similarly, Justin Kinney has been a source of inspiration and help, and an all around fun guy to work with.

I'd like to thank Brett Lomenick, Annie Moradian, Mike Sweredoski, and everyone at the PEL for helping me through some very tough experiments and always being available to give me advice.

Lastly I'd like to thank my parents for always supporting me and believing in me.

ABSTRACT

Measurements in biology have reached a level of precision that demands quantitative modeling. This is particularly true in the field of gene regulation, where concepts from physics such as thermodynamics have allowed for accurate models to be made.

Many issues remain. DNA sequencing is routine enough to sequence new genomes in days and cheap enough to use deep sequencing to perform precision measurements, but our ability to interpret the wealth of genomic data is lagging behind, especially in the realm of gene regulation. The primary reason is that we lack any information what so ever as to the basic regulatory details of ≈ 65 percent of operons even in *E. coli*, the best understood organism in biology. As a result we cannot use our hard won modeling efforts to understand any of these operons.

This work takes steps to address these issues. First we use 30 LacI mutants as a test case to prove that we can make quantitatively accurate models of gene expression and sequence-dependent binding energies of transcription factors and RNA polymerase.

Next we note that much of the quantitative insight available on transcriptional regulation relies on work on only a few model regulatory systems such as LacI as was considered above. We develop an approach, through a combination of massively parallel reporter assays, mass spectrometry, and information-theoretic modeling that can be used to dissect bacterial promoters in a systematic and scalable way. We demonstrate that we can uncover a qualitative list of transcription factor binding sites as well as their associated quantitative details from both well-studied and previously uncharacterized promoters in *E. coli*.

Finally we extend the above method to over 100 *E. coli* promoters using over 12 growth conditions. We show the method recapitulates known regulatory information. Then, we examine regulatory architectures for more than 80 promoters which previously had no known regulation. In many cases, we identify which transcription factors mediate their regulation. The method introduced clears a path for fully characterizing the regulatory genome of *E. coli* and advances towards the goal of using this method on a wide variety of other organisms including other prokaryotes and eukaryotes such as *Drosophila melanogaster*.

PUBLISHED CONTENT AND CONTRIBUTIONS

Ireland, W. T. et al. (2020). “Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time.” In: *bioRxiv*. DOI: 10.1101/2020.01.18.910323.

W. T. I. helped conceive of the project, analyzed results, prepared data, and co-wrote the manuscript.

Barnes, Stephanie L. et al. (2019). “Mapping DNA sequence to transcription factor binding energy *in vivo*”. In: *PLoS Computational Biology* 15.2, pp. 1–29. DOI: 10.1371/journal.pcbi.1006226.

W. T. I. helped conceive of the project, analyzed results, prepared data, and co-wrote the manuscript.

Belliveau, Nathan M. et al. (2018). “Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.21, E4796–E4805. DOI: 10.1073/pnas.1722055115.

W. T. I. helped conceive of the project, analyzed results, prepared data, and co-wrote the manuscript.

Ireland, William T. and Kinney (2016). “MPAthic: Quantitative Modeling of Sequence-Function Relationships for massively parallel assays”. en. In: DOI: 10.1101/054676.

W. T. I. helped conceive of the project, wrote analysis code, and co-wrote the manuscript.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	iv
Published Content and Contributions	v
Table of Contents	v
List of Illustrations	viii
List of Tables	xi
Chapter I: Introduction	1
1.1 Chapter Summaries	2
1.2 The central dogma of molecular biology	5
1.3 Thermodynamic models and gene regulation	7
1.4 Quantitative modeling of gene regulation	15
1.5 Lack of current regulatory knowledge	16
Chapter II: Mapping DNA Sequence to Transcription Factor Binding Energy	
<i>in vivo</i>	23
2.1 Introduction	23
2.2 Results	25
2.3 Discussion	40
2.4 Methods	42
Chapter III: A Systematic Approach for Dissecting the Molecular Mechanisms	
of Transcriptional Regulation in Bacteria.	52
3.1 Introduction	52
3.2 Results	54
3.3 Discussion	73
3.4 Methods	74
3.5 Bacterial strains	76
3.6 Supplemental Information: Characterization of library diversity and	
sorting sensitivity.	84
3.7 Analysis of library diversity using data from the <i>mar</i> promoter.	84
3.8 Supplemental Information: Generation of sequence logos.	85
3.9 Generating position weight matrices from known genomic binding	
sites.	85
3.10 Generating position weight matrices from Sort-Seq data.	87
3.11 Supplementary Information: Additional data and analysis for <i>yebG</i> ,	
<i>purT</i> , <i>xylE</i> , and <i>dgoR</i>	87
Chapter IV: Deciphering the regulatory genome of <i>Escherichia coli</i> , one	
hundred promoters at a time	100
4.1 Introduction	100
4.2 Results	104
4.3 Discussion	119

4.4 Methods	122
Appendix A: Extended experimental details	140
Appendix B: Extended analysis details	146
B.1 Validating Reg-Seq against previous methods and results	146
B.2 Information footprints	150
B.3 Estimating mutual information from observed data and model pre- dictions	154
B.4 Markov Chain Monte Carlo fitting procedure	155
B.5 TOMTOM motif comparison	159
B.6 BioInformatic methods - TOMTOM motif comparison	160
B.7 Sigma Factors	162
B.8 Binding sites regulating divergent operons	163
B.9 Neural network fitting	166
B.10Diffeomorphic modes	169
B.11Diffeomorphic Mode Calculations	174
B.12Genes	177
B.13Construction of sequence logos	179

LIST OF ILLUSTRATIONS

<i>Number</i>	<i>Page</i>
1.1 Applying quantitative models of gene regulation across the genome. . .	3
1.2 The central dogma of molecular biology.	7
1.3 Input-output function for simple repression.	8
1.4 Modeling transcription using statistical mechanics.	9
1.5 States and weights for a simple repression architecture.	10
1.6 States and weights for a simple activation architecture.	11
1.7 Identification of operons in <i>E. coli</i> with and without regulatory an- notation.	16
1.8 Distribution of regulatory architectures in <i>E. coli</i>	17
1.9 Proteome Data from Schmidt et al., 2016.	19
2.1 Process flow for using Sort-Seq to obtain energy matrices.	26
2.2 States and weights for the simple repression motif.	28
2.3 Inference of LacI energy matrices.	31
2.4 Energy matrices for the natural lac operators from Sort-Seq data. . . .	32
2.5 Fold-change data reflects expected values from predicted fold change curves.	35
2.6 Energy matrix predictions can be used to design precise phenotypic responses.	37
2.7 Point mutations to LacI DNA-binding domain cause subtle changes to sequence specificity.	40
3.1 Summary of transcriptional regulatory knowledge in <i>E. coli</i>	53
3.2 Overview of approach to characterize transcriptional regulatory DNA, using Sort-Seq and mass spectrometry.	56
3.3 Sort-Seq identifies the regulatory landscape of the <i>lac</i> , <i>rel</i> , and <i>mar</i> promoters.	58
3.4 Comparison between Sort-Seq and genomic-based sequence logos. . .	60
3.5 DNA affinity purification and identification of LacI and RelBE by mass spectrometry using known target binding sites.	61
3.6 Identification of transcription factors using DNA-affinity chromatog- raphy and mass spectrometry.	63

3.7	Identification of unannotated genes with potential regulation and distribution of known transcription factor binding sites in <i>E. coli</i>	65
3.8	Sort-Seq distinguishes directional regulatory features and uncovers the regulatory architecture of the <i>purT</i> promoter.	66
3.9	Sort-Seq identifies a set of activator binding sites that drive expression of RNAP at the <i>xylE</i> promoter.	68
3.10	<i>lexA</i> and <i>yebG</i> regulation.	70
3.11	The <i>dgoRKADT</i> promoter is induced in the presence of D-galactonate due to loss of repression by DgoR and activation by CRP.	72
3.12	Related to Fig. 3.2 and Fig. 3.3. Analysis of the library mutation spectrum and effect of Sort-Seq sorting conditions.	79
3.13	Extended analysis of the <i>dgoR</i> promoter.	90
4.1	The <i>E. coli</i> regulatory genome and the genes studied with Reg-Seq.	105
4.2	The Reg-Seq procedure used to determine how a given promoter is regulated.	106
4.3	A summary of regulatory architectures discovered in this study.	112
4.4	Newly discovered or updated regulatory architectures.	114
4.5	Examples of the insight gained by Reg-Seq in the context of promoters with no previously known regulatory information.	115
4.6	Reg-Seq analysis of broadly-acting transcription factors.	131
4.7	Inspection of an anaerobic respiration genetic circuit.	132
4.8	Representative view of the interactive figure that is available online.	132
4.9	All regulatory cartoons for genes considered in Reg-Seq.	133
4.10	The p-value distribution from TOMTOM for comparisons to FNR and GlpR binding sites.	134
A.1	Promoter constructs for Reg-Seq.	142
B.1	A summary of four direct comparisons of measurements using fluorescence and sorting and using RNA-Seq.	147
B.2	Reg-Seq analysis of “gold standard” promoters.	149
B.3	Parameter inference using Markov Chain Monte Carlo.	156
B.4	An example energy matrix for the YieP binding site of the <i>ykgE</i>	158
B.5	A comparison of RNAP -10 site sequence logos.	160
B.6	Motif comparison using TOMTOM.	161
B.7	Sequence logos for neural network methods.	162
B.8	Multipurpose binding sites.	163
B.9	Comparison of Reg-Seq architectures to RegulonDB.	167

B.10	Architecture of neural network used to fit data.	169
B.11	The microstates of a one transcription factor promoter.	170
B.12	A sequence logo of the <i>tff</i> which has one RNAP binding site and is repressed by GlpR.	170
B.13	Data from Reg-Seq	210

LIST OF TABLES

<i>Number</i>	<i>Page</i>
2.1 Summary of LacI binding site mutant energy prediction for designed O1 sites.	34
2.2 Summary of all energy predictions for mutant constructs.	46
4.1 All promoters examined in Reg-Seq, categorized according to type of regulatory architecture.	113
4.2 All genes investigated in this study categorized according to their regulatory architecture	130
A.1 All TSS for all genes investigated in Reg-Seq.	145
B.1 All consensus σ factor binding sites	163
B.2 Identification of the σ factors used for each RNAP binding site. . . .	166

Chapter 1

INTRODUCTION

We live in the "genomic era" where DNA sequencing is routine enough to sequence new genomes in days and deep sequencing is used to get precision measurements. One of the myriad examples of these measurements comes from measuring ribosome occupancy (Ingolia et al., 2009), where measurements are taken of the probability that a ribosome, the key piece of machinery for producing new proteins, is bound to its target sequence. As the number of ribosomes bound will be directly proportional to protein produced, this provides a useful measure of how much of that protein exists in the cell. Another application of sequencing technology is in measuring gene expression directly (Melnikov et al., 2012). Cheap sequencing has led to the development of over one hundred sequencing-based methods (Pachter, 2013).

We also live in a growing era of quantitative biology. Measurements in biology are growing increasingly precise. Massively parallel reporter assays (MPRAs) can use hundreds of thousands of designed DNA constructs to make measurements (Kinney and McCandlish, 2019) where a few decades ago creating and testing a single piece of mutant DNA inside cells would be a project in and of itself. As the ability to assess a huge number of perturbations has revolutionized the inputs to quantitative biology experiments, RNA-seq combined with deep sequencing revolutionized measurement of the outputs. RNA-seq can measure the gene expression of those hundreds of thousands of designed promoter constructs in parallel. Super resolution microscopy techniques can measure *in vivo* protein dynamics of objects on the order of nanometers (Cisse et al., 2013), and the interactions between transcription factors and RNAP, both crucial factors for gene regulation, can be measured on the order of thousandths of an eV (Forcier et al., 2018). Furthermore, there are phenomenological findings such as phase separation contributing to gene regulation in eukaryotes (Cisse, 2020). Such measurements demand commensurate theory, a theory which in large part still lags behind. Ideas borrowed from physics have contributed greatly to theory in biology. Theory on gene regulation in particular has benefited greatly from the power of statistical mechanical thinking in the biological setting, and in this dissertation we discuss our efforts to extend our ability to quantitatively model gene regulation throughout *E. coli*. We validate a method for recovering a base pair resolution map of gene regulation in *E. coli* and develop it into a high throughput

tool that in the future can be utilized on other organisms to provide the quantitative data and qualitative details necessary to unravel the continuing mystery of gene regulation.

1.1 Chapter Summaries

In Chapter 1 we discuss the necessary background to understand gene regulation, namely the central dogma of biology and the action of transcription factors and RNA polymerase (RNAP). We derive some of the models from statistical mechanics we use to model gene expression, and discuss the woeful lack of basic regulatory knowledge in *E. coli* that hinders our efforts to understand gene regulation and completely stops us from applying quantitative modeling across the larger *E. coli* genome.

In Chapter 2 we discuss the modeling of DNA sequence-specific transcription factor binding energies *in vivo*. We create models that allow us to predict the binding energy between a transcription factor and a mutated version of its binding site using Lac repressor as a test bed. We demonstrate our ability to generate accurate models by comparing model predictions from Sort-Seq to independent measurements of DNA transcription factor interactions using microscopy. We then show that this modeling technique can be used to address a number of scientific questions. For example, we observe how the preferred DNA sequence for transcription factor binding changes when amino acid mutations are made to the transcription factor's DNA binding domain, which helps us to understand how transcription factors and their binding sites co-evolve. This provides yet another example of the importance of quantitative models for deeply understanding biological mechanisms. A summary of what we will discuss is displayed in Fig 1.1(A).

In Chapter 3 we acknowledge that despite an ability to build models based on statistical mechanics for gene regulation, these models cannot be used in the vast majority of cases. This is because to build models we must first know some crucial details of the regulatory context. Specifically, to even begin to make predictions we need to know what transcription factors bind to the DNA for a given operon. Even for *E. coli*, the best understood organism in biology, we know nothing about the regulatory details for $\approx 65\%$ of operons (Gama-Castro et al., 2016). Past efforts to understand regulatory sequences and solve the regulatory ignorance problem on a large scale have failed to solve the problem. One such method is to use computational methods to "sequence gaze", or in other words, to search the *E. coli*

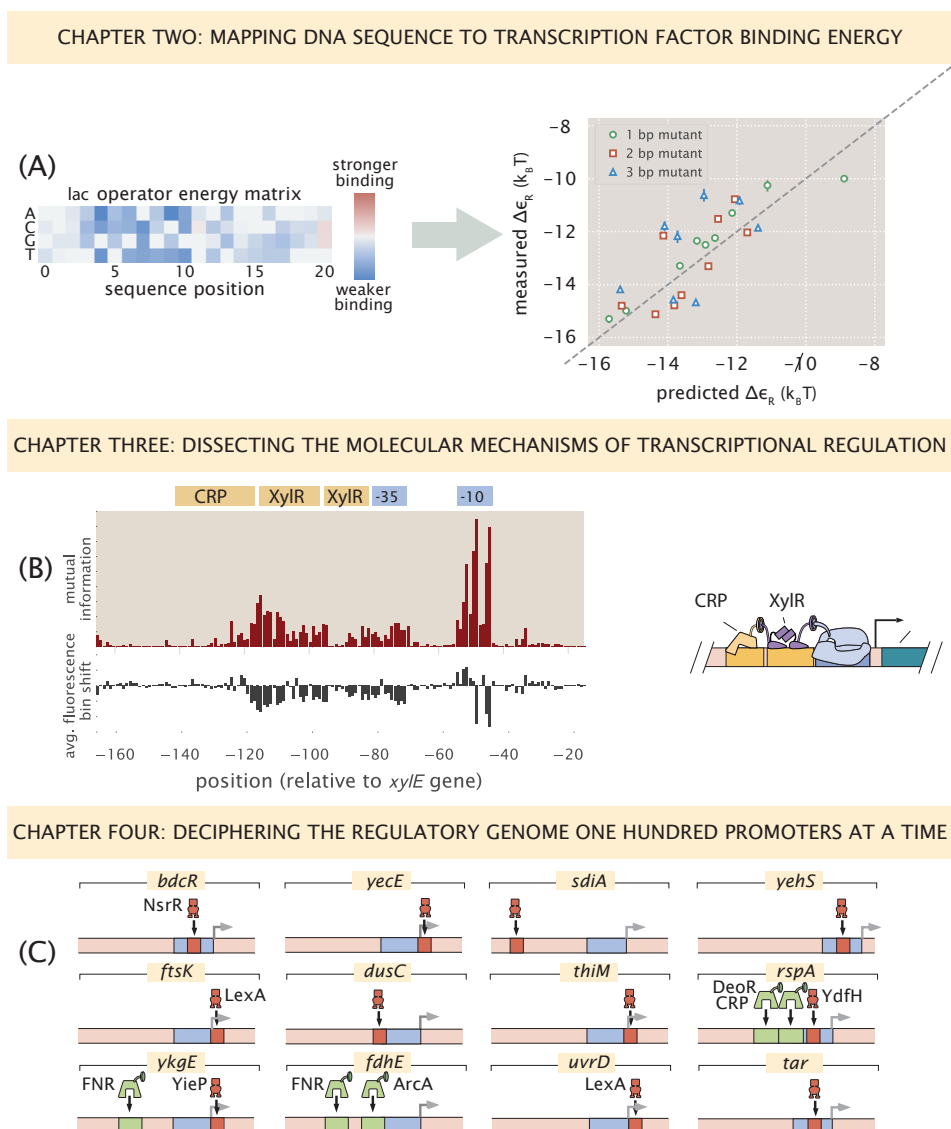


Figure 1.1: Applying quantitative models of gene regulation across the genome. (A) In Chapter 2 we used *in vivo* techniques to infer energy matrices in absolute energy units (k_bT). We used these energy matrices to predict the binding energies of Lac O1 binding site mutants and confirmed our predictions with experimental measurements (right). (B) In Chapter 3 we identified regulatory architectures for unannotated promoters. We quantified the mutual information between gene expression and mutation at each sequence position in the promoter (left) and combined these observations with DNA affinity chromatography and mass spectrometry to infer regulatory architectures (right). (C) In Chapter 4 we take the success of Sort-Seq and adapt it for use across orders of magnitude more genes. We go over the regulatory elements from the over 100 *E. coli* promoters studied.

genome for sequences that are similar to known examples of transcription factor binding sites (Compan and Touati, 1994; Kumar and Shimizu, 2011; Easton and Kushner, 1983). As any computational method searches through 4.6 megabases of DNA, and therefore an equal number of possible binding sites, it is unsurprising that these methods tend to yield some false positives. We disprove several of these computationally discovered binding sites in Chapter 4. Other methods for discovering binding sites on a wide scale such as chromatin immunoprecipitation (Bonocora and Wade, 2015) do not provide base pair resolution and cannot determine how transcription factors interact with RNAP or each other. Lastly, while *in vitro* methods such as protein-binding microarrays (Berger et al., 2006), SELEX (Fields et al., 1997; Jolma et al., 2013) and MITOMI (Maerkl and Quake, 2007; Shultzaberger et al., 2012) can provide useful insights, they can never fully account for *in vivo* effects.

To tackle the regulatory ignorance problem and get a base pair resolution picture of regulation, we apply Sort-Seq (Kinney and Callan, 2010), to characterize the regulatory DNA. We further develop the method as a way to systematically approach the regulation of any promoter quantitatively. Here we first apply Sort-Seq across 6 different bacterial promoters to uncover the functional binding sites where transcription factors bind to regulate gene expression. Using DNA affinity chromatography and mass spectrometry we then identify the transcription factors that bind these sites, and apply information-theoretic modeling to infer energy matrix models of binding by each transcription factor. We validate the approach by applying it to the well-characterized promoters of *lacZYA*, *relBE*, and *marRAB*. We then demonstrate that it can work equally well to uncover the previously uncharacterized regulatory architectures for the promoters of *purT*, *xylE*, and *dgoRKADT*. A summary of what we will discuss is displayed in Fig 1.1(B).

In Chapter 4 we take the success of the Sort-Seq methodology and scale up by an order of magnitude to show that it can be applied across the genome. In Sort-Seq, only one gene at a time could be investigated, which made it extremely difficult to use the method on a wide scale. One bottleneck in the process was our measurement method itself. While using fluorescence based cell sorting (FACS) on a single gene was a short process, it was not readily parallelizable, and when trying to tackle even tens of operons under multiple growth conditions, the sorting time alone would make it difficult to carry out the experiment. We transition the fluorescence based measurement methodology of Sort-Seq to an RNA-seq based measurement

methodology. Using RNA-seq as a measurement tool we were able to measure expression for 100 genes of interest simultaneously, and there is no limit to scaling up to measuring the expression of every operon in *E. coli* simultaneously.

In Chapter 4, we discuss how we produced a base pair resolution dissection of more than 100 *E. coli* promoters in 12 growth conditions. We show the method recapitulates known regulatory information. Specifically we once again examine several of the genes investigated using Sort-Seq, namely *lacZYA*, *relBE*, *marRAB*, and *dgoRKADT*. The correspondence is demonstrated in Fig B.2. Then, we examine regulatory architectures for more than 80 promoters which previously had no known regulation. In many cases, we identify which transcription factors mediate their regulation. An summary of what we will discuss is displayed in Fig 1.1(C). Techniques in DNA-synthesis and microbiology are becoming sufficient to use Reg-Seq throughout *E. coli* and also on other organisms such as *Drosophila* or *Pseudomonas aeruginosa*. Not only could these new systems eventually become model organisms in their own right, but elucidating regulatory details in eukaryotic systems is one of the necessary steps in extending the modeling success of prokaryotes to eukaryotes.

1.2 The central dogma of molecular biology

The hard won knowledge of the genetic code has been the greatest accomplishment of molecular microbiology. We can see in Fig. 1.2 the "central dogma" of molecular biology.

To translate DNA, the hereditary material of the cell, into a proteins, which perform most of the useful tasks in the cell, the protein coding region, known as a "gene", must first be copied into a message that can be read by the ribosomes which then build the proteins. In a process known as transcription, an RNA polymerase (RNAP) recognizes and binds to a region upstream of the gene known as a promoter, and then copies the gene into a single-stranded RNA message known as mRNA. Next, the mRNA is read by a ribosome in a process known as translation to produce the final protein.

The remarkable thing about this process is that it is conserved throughout all organisms, even those as distantly related as bacteria and vertebrates, earning it the title of "central dogma."

It was the culmination of a decades-long search to unravel the mechanism by which genetic information is passed down from generation to generation in living organisms and then to discover how the steps of the central dogma are carried out to produce

useful products.

Oswald Avery discovered that DNA (and not protein) is the molecule by which genetic information is propagated (Avery, MacLeod, and McCarty, 1944). Watson and Crick discovered the helical structure of DNA, which immediately suggested a possible copying mechanism for the genetic material (Watson and Crick, 1953). Crick, Brenner, and coworkers arrived at the now familiar result that a protein coding sequence consists of a series of trinucleotide codons (F. Crick, 1961). Subsequent work was able to provide a codon table that can translate any three base pair sequence into a corresponding amino acid. As a result, we have a deep understanding of the protein coding regions of the genetic code.

However, in many ways the central dogma remains a mystery. The non-coding regions of the genome have no such corresponding mapping. These regulatory regions control the levels of protein expression and are important to how organisms respond to the environment and are crucial for the fitness of the organism. The functions of the regulatory regions are as much of a mystery as they were decades ago. For an arbitrary DNA sequence in a regulatory region we have no knowledge whatsoever as to its function. While there are several ways in which protein copy number can be controlled, this dissertation focuses on how the DNA sequence of the regulatory region controls how DNA is transcribed into mRNA, called transcriptional regulation.

In general transcriptional regulation is accomplished by modulating the probability that RNAP will bind to the promoter and proceed to copy the gene, which is known as the occupancy hypothesis (Ackers and Johnson, 1982). The probability of RNAP binding depends in part on the sequence of the promoter itself, as the polymerase has DNA sequence binding preferences and deviating from these preferences will reduce the probability of binding. The DNA binding preference of a protein is often displayed as a "consensus sequence". A consensus sequence is the ideal series of nucleotides for protein binding, generally calculated by looking across all binding sites in the genome and finding the most common base pair at each position in the binding site. For example, the consensus sequence of the RNAP -10 region is $_{-15}\text{TGNTATAAT}_{-7}$, where N represents having no nucleotide preference at that site and the numbers -15 and -7 represent locations of that base pair as compared to the transcription start site (TSS).

However, the promoter sequence is static and cannot respond to changes in environment or growth state. In order to enact transcriptional regulation that can change

in response to an external stimulus, the cell produces DNA-binding proteins known as transcription factors that bind to the promoter near the RNAP binding site and control RNAP binding probability. For example, transcription factors known as repressors often will bind to DNA near the RNAP site. This will physically occlude the RNAP from binding, decreasing the RNAP occupancy, and therefore decreasing protein production. Similarly, activators are DNA-binding proteins that interact with RNAP, forming favorable energetic interactions, and therefore making it more likely for RNAP to be bound to its binding site.

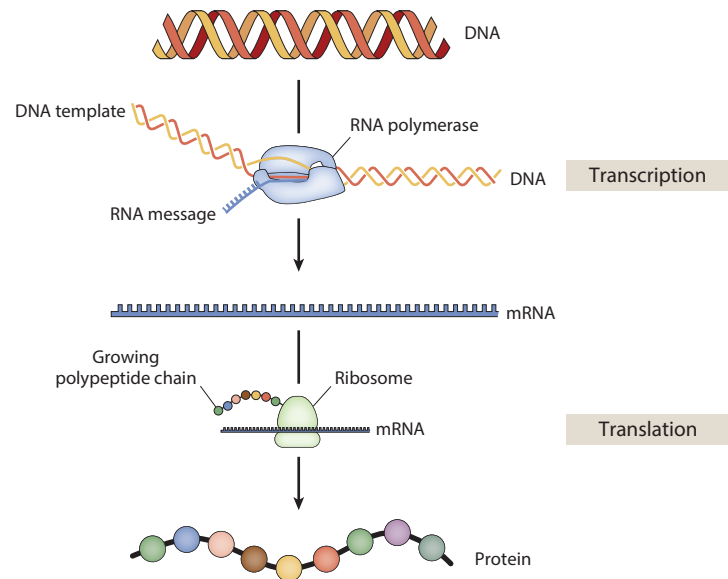


Figure 1.2: The central dogma of molecular biology. Genes are encoded as DNA sequences within the genome. RNA polymerase (RNAP) copies the DNA as a single-stranded mRNA transcript. Then, ribosomes translate the mRNA into protein by facilitating the pairing of tRNAs with the mRNA transcript and joining the associated amino acids together into a polypeptide chain. This polypeptide chain then generally self-assembles into a protein.

1.3 Thermodynamic models and gene regulation

A primary tool that has been borrowed from physics to quantify gene expression is the use of equilibrium statistical mechanics. While life is one of the most interesting examples of a dynamic, out of equilibrium system, gene regulation is one of many examples in biology in which equilibrium formulations of ideas from physics have surprising utility (Phillips, 2015). We see in Fig. 1.3 that these classes of models have allowed us to quantitatively predict output (gene expression) over several orders of magnitude in input (transcription factor copy number).

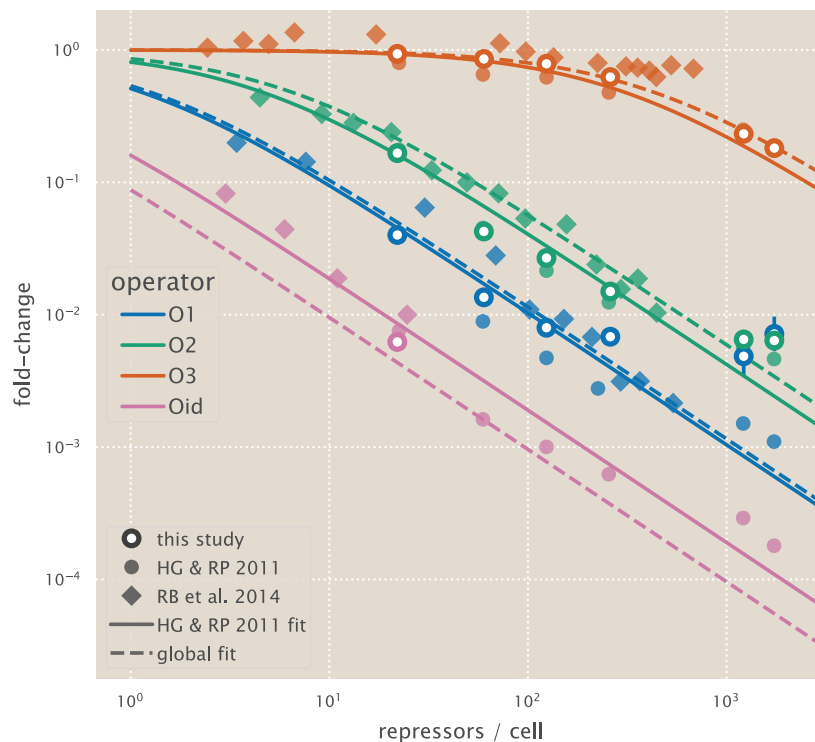


Figure 1.3: Input-output function for simple repression. For a simple repression architecture, with one RNAP site, and one repressor binding site, Garcia and Phillips, 2011 measured the output of *lacZ* for transcription factor binding sites of different DNA-protein binding strength. Figure data taken from Garcia and Phillips, 2011

Statistical mechanics concerns itself with the probability of different microstates in systems containing a large number of interacting particles. A microstate is a unique arrangement of particles, which may or may not have properties that are distinguishable from other microstates. The probability of a specific microstate is given by the Boltzmann distribution,

$$p(\varepsilon_i) = \frac{1}{Z} e^{-\beta \varepsilon_i}, \quad (1.1)$$

where ε_i is the energy of microstate i . Z is the partition function, and β is equal to $1 = k_b T$ where k_b is the Boltzmann constant and T is the temperature of the system. The quantity $e^{-\beta \varepsilon_i}$ is referred to as the Boltzmann factor. The partition function can be thought of the sum of the statistical mechanical weights of all microstates in the system, and is given by

$$Z = \sum_{i=1}^N e^{-\beta \varepsilon_i}. \quad (1.2)$$

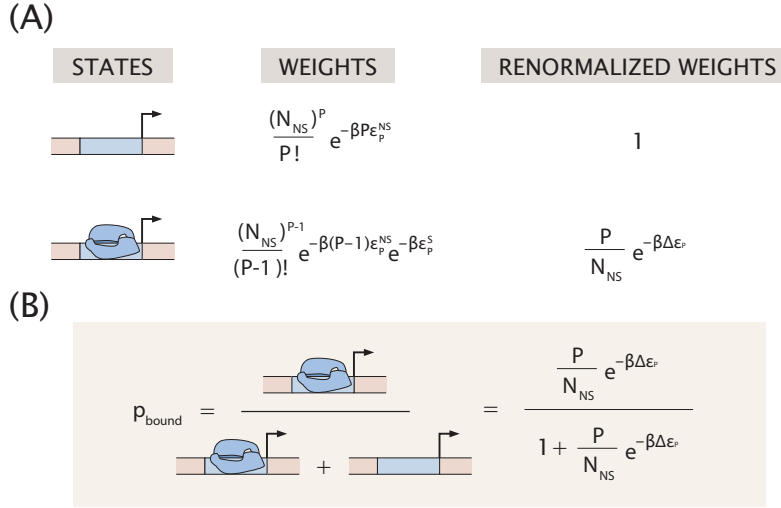


Figure 1.4: Modeling transcription using statistical mechanics. To model gene expression, we make the assumption that gene expression is proportional to the probability that RNAP is bound to the promoter, p_{bound} (Ackers and Johnson, 1982). (A) To determine the value of p_{bound} we then enumerate all of the states available to the system and assign statistical mechanical weights based on the energy associated with each state and the multiplicity of each state. Renormalizing the weights such that the unbound state has a weight of 1 then provides us with a clean set of statistical mechanical weights that can be used to determine the value of p_{bound} . (B) The value of p_{bound} is equal to the statistical mechanical weight of the RNAP bound state divided by the sum of the weights of all possible states.

When modeling transcription, our goal is to determine the probability that an RNAP will bind to a promoter and initiate transcription. We assume that RNAP occupancy is proportional to total gene expression, an assumption known as the occupancy hypothesis. When using a statistical mechanical approach, we identify the various states that a system can adopt, where a state is a set of microstates with indistinguishable properties. We assign statistical mechanical weights to each state and use these weights to determine the probability of RNAP binding, p_{bound} . This identification of states and weights is modeled for the case of constitutive transcription in Fig. 1.4 (A).

Any combination of regulatory proteins and RNAP can be modeled using statistical mechanics. Next, we provide a derivation of a statistical mechanical expression

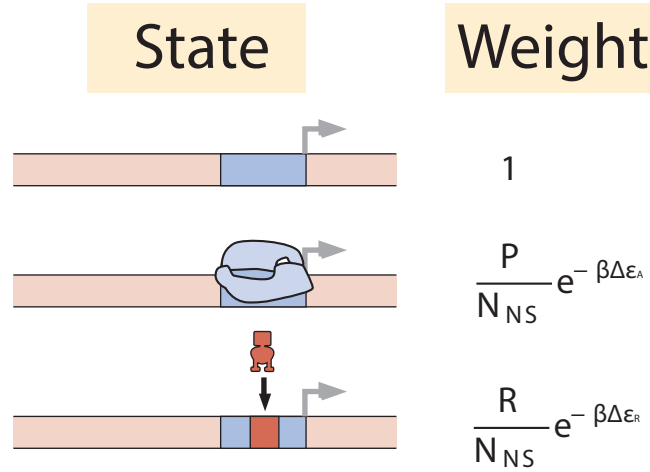


Figure 1.5: States and weights for a simple repression architecture. Simple repression occurs when a single transcription factor binds in the vicinity of the RNAP binding site and prevents RNAP binding.

for the probability of RNAP binding at a constitutive promoter. We show how this derivation can be represented by a states and weights diagram, an approach which can then be generalized to more complex regulatory scenarios.

In the case of constitutive transcription, there are many copies of RNAP and many DNA binding sites available to the RNAP. A microstate can be thought of as a “snapshot” of the positions of all RNAP relative to the genome at a given time. If we are interested in the transcription of a specific gene, then we wish to know the probability that a single copy of RNAP is bound to that gene’s promoter. We can determine this probability using Equation 1.1 provided we know the energy ϵ_i of the state, the multiplicity of the state (i.e., the number of possible microstates in which RNAP is bound to the promoter of interest) and the partition function Z that represents all possible microstates of the system.

To simplify the problem, we abstract the genome as a single specific RNAP binding site and a series of nonspecific binding sites that bind weakly with the RNAP. In reality, there are many specific RNAP binding sites in the genome with a distribution of strengths, and 10% of 100 bp regions have at least one active RNAP site (Yona, Alm, and Gore, 2018). For the purpose of this problem, we will view DNA aside from our binding site of interest as being part of a “pool” of DNA binding sites with some average weak binding energy. There are N_{NS} of these nonspecific binding

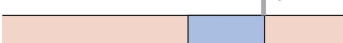
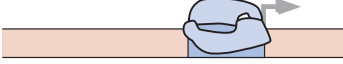
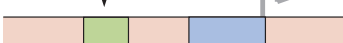

State	Weight
	1
	$\frac{P}{N_{NS}} e^{-\beta \Delta \epsilon_P}$
	$\frac{A}{N_{NS}} e^{-\beta \Delta \epsilon_A}$
	$\frac{PA}{N_{NS}} e^{-\beta (\Delta \epsilon_A + \Delta \epsilon_P + \epsilon_{AP})}$

Figure 1.6: States and weights for a simple activation architecture. Simple activation occurs when a single transcription factor binds in the vicinity of the RNAP binding site and promotes RNAP binding.

sites, where we assume that N_{NS} is approximately equal to the length of the genome. We define ϵ_P^S as the energy of an RNAP bound to the specific binding site and ϵ_P^{NS} to an RNAP bound to any of the nonspecific sites.

The energy of any microstate i in which an RNAP is bound to the specific site must account for both the energy of one RNAP binding to the specific site and $P - 1$ RNAPs binding to nonspecific sites, where P is the total number of RNAPs in the system, such that $\epsilon_i = (P - 1)\epsilon_P^{NS} + \epsilon_P^S$. The Boltzmann factor for such a microstate is thus $e^{-\beta(P-1)\epsilon_P^{NS}} e^{-\beta\epsilon_P^S}$. The value of p_{bound} is given by the sum of the Boltzmann weights for all microstates in which an RNAP is bound to the specific site, giving us

$$p_{bound} = \frac{\sum_{i=1}^N e^{-\beta(P-1)\epsilon_P^{NS}} e^{-\beta\epsilon_P^S}}{Z_{tot}}, \quad (1.3)$$

where we define Z_{tot} as the total partition function. We can rewrite p_{bound} as

$$p_{bound} = \frac{e^{-\beta\epsilon_P^S} Z_{NS}(P - 1, N_{NS})}{Z_{tot}}, \quad (1.4)$$

where $Z_{NS}(P - 1, N_{NS})$ is a partial partition function representing all microstates in which $P - 1$ RNAP are distributed among N_{NS} nonspecific binding sites, as will

occur when one RNAP is specifically bound. We can further define Z_{tot} as

$$Z_{tot} = e^{-\beta \varepsilon_P^S} Z_{NS}(P-1, N_{NS}) + Z_{NS}(P, N_{NS}), \quad (1.5)$$

which gives us

$$p_{bound} = \frac{e^{-\beta \varepsilon_P^S} Z_{NS}(P-1, N_{NS})}{e^{-\beta \varepsilon_P^S} Z_{NS}(P-1, N_{NS}) + Z_{NS}(P, N_{NS})}. \quad (1.6)$$

We can now see that the Equation for p_{bound} is of the form a Boltzmann distribution where the states are either RNAP bound, which consists of all microstates in which an RNAP is bound to the specific site and has a weight given by $e^{-\beta \varepsilon_P^S} Z_{NS}(P-1, N_{NS})$ or RNAP unbound, which consists of all microstates in which no RNAP is bound to the specific site and has a weight given by $Z_{NS}(P, N_{NS})$. A illustration of these states is shown in the states column of Fig. 1.4.

Next we wish to rewrite Equation 1.6 using measurable parameters. A partition function can be thought of as the product of a state's Boltzmann factor and the state's multiplicity, or the number of microstates where, for example, RNAP is bound. We have already determined the Boltzmann factors for each state in our model, and the multiplicities can be determined combinatorially. Doing so gives us the statistical mechanical weight of the bound state,

$$e^{-\beta \varepsilon_P^S} Z_{NS}(P-1, N_{NS}) = \frac{(N_{NS})!}{(P-1)!(N_{NS}-P+1)!} e^{-\beta(P-1)\varepsilon_P^{NS}} e^{-\beta \varepsilon_P^S} \quad (1.7)$$

and the weight of the unbound state,

$$Z_{NS}(P, N_{NS}) = \frac{(N_{NS})!}{(P)!(N_{NS}-P)!} e^{-\beta P \varepsilon_P^{NS}}. \quad (1.8)$$

These weights can be simplified using the approximation

$$\frac{(N_{NS})!}{P!(N_{NS}-P)!} \approx \frac{(N_{NS})^P}{P!}, \quad (1.9)$$

where $N_{NS} \gg P$. The simplified weights are represented in the weights column of Fig. 1.4. We can write p_{bound} as

$$p_{bound} = \frac{\frac{(N_{NS})^{P-1}}{(P-1)!} e^{-\beta(P-1)\varepsilon_P^{NS}} e^{-\beta\varepsilon_P^S}}{\frac{(N_{NS})^{P-1}}{(P-1)!} e^{-\beta(P-1)\varepsilon_P^{NS}} e^{-\beta\varepsilon_P^S} + \frac{(N_{NS})^P}{(P)!} e^{-\beta P\varepsilon_P^{NS}}}. \quad (1.10)$$

Finally, we can greatly simplify the form of the equation by dividing the weight for each state by the weight of the unbound state. The unbound state then has a renormalized weight equal to 1, and the bound state has a renormalized weight of $\frac{P}{N_{NS}} e^{-\beta(\varepsilon_P^S - \varepsilon_P^{NS})}$. We define $\Delta\varepsilon_P = \varepsilon_P^S - \varepsilon_P^{NS}$ where $\Delta\varepsilon_P$ represents the difference in RNAP binding energy between the specific binding site and the nonspecific genomic background. The renormalized weights for each state are illustrated in Fig. 1.4 in the renormalized weights column. Substituting the renormalized values into Eq. 1.10 gives us

$$p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}. \quad (1.11)$$

The process of deriving the equation for p_{bound} is identical to that used for more complex regulatory scenarios. Results for many architectures are considered in Bintu et al., 2005, and we now consider the cases of simple activation and simple repression in detail.

Simple Repression

We consider the case of simple repression, in which a repressor binds adjacent to an RNAP binding site and prevents RNAP from binding. In this case there are three states available to the system: no proteins bound, repressor bound, and RNAP bound. These states and their associated weights are displayed in Fig. 1.5. The expression for the probability of RNAP binding, p_{bound} in a simple repression architecture is found in a manner identical to that of the constitutive expression scenario, namely we divide the statistical weight of all states with RNAP bound by the total partition function, which yields

$$p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P} + \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_R}}. \quad (1.12)$$

As noted previously, it is assumed that gene expression is proportional to p_{bound} . However, it is difficult to determine the exact proportionality between these quantities, and we lack a straightforward way to measure p_{bound} *in vivo* in order to fix

unknown parameters. It is therefore more convenient to think about gene regulation using fold-change. For a constitutive promoter, the fold change is shown in Fig. 1.4 (B). Fold-change quantifies the change in expression due to regulation. This quantity is straightforward to measure experimentally and has a clear interpretation in regards to regulatory strength. For repression the fold-change is given by

$$\text{fold-change} = \frac{p_{\text{bound}}(R)}{p_{\text{bound}}(R=0)}. \quad (1.13)$$

We can substitute Eq. 1.12 into Eq. 1.13, which gives us

$$\text{fold-change} = \left(\frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P} + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_R}} \right) \left(\frac{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}} \right). \quad (1.14)$$

To simplify this expression, we make use of the weak promoter approximation, where we assume RNAP binds weakly to the promoter which implies that $\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P} \ll 1$. We can then simplify to

$$\text{fold-change} \approx \frac{1}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_R}}. \quad (1.15)$$

Simple Activation

The case of simple activation is similar to simple repression, though it incorporates the additional factor of cooperative interactions between proteins (namely the RNAP and the activator). In simple activation, an activator and RNAP can bind to the promoter simultaneously, as noted in the states and weights diagram for simple activation shown in Fig. 1.6. The binding of multiple proteins gives this state a multiplicity of $\frac{A}{N_{NS}} \frac{P}{N_{NS}}$, where A is the number of activators in the system. An interaction energy between the activator and RNAP, ε_{ap} must be included in the Boltzmann factor and for activators is always a favorable interaction which will serve to make the doubly bound state more likely. A typical value for ε_{ap} is $\approx -4k_bT$ which is then represented as $e^{-\beta(\Delta \varepsilon_A + \Delta \varepsilon_P + \varepsilon_{ap})}$ where $\Delta \varepsilon_A$ represents the binding energy of the activator to its binding site.

Given these adjustments, p_{bound} can then be written as

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P} + \frac{A}{N_{NS}} \frac{P}{N_{NS}} e^{-\beta(\Delta \varepsilon_A + \Delta \varepsilon_P + \varepsilon_{ap})}}{1 + \frac{A}{N_{NS}} e^{-\beta \Delta \varepsilon_A} + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P} + \frac{A}{N_{NS}} \frac{P}{N_{NS}} e^{-\beta(\Delta \varepsilon_A + \Delta \varepsilon_P + \varepsilon_{ap})}}. \quad (1.16)$$

We can use the weak promoter approximations $\frac{P}{N_{NS}}e^{-\beta\Delta\epsilon_P} \ll 1$ and $\frac{P}{N_{NS}}e^{-\beta(\Delta\epsilon_P+\epsilon_{ap})} \ll 1$ to simplify to

$$p_{bound} \approx \frac{\frac{P}{N_{NS}}e^{-\beta\Delta\epsilon_P} + \frac{A}{N_{NS}}\frac{P}{N_{NS}}e^{-\beta(\Delta\epsilon_A+\Delta\epsilon_P+\epsilon_{ap})}}{1 + \frac{A}{N_{NS}}e^{-\beta\Delta\epsilon_A}}. \quad (1.17)$$

As in Eq. 1.13, the fold-change for the activator can be written as

$$\text{fold-change} = \frac{p_{bound}(A)}{p_{bound}(A=0)}, \quad (1.18)$$

and then simplified to

$$\text{fold-change} \approx \frac{1 + \frac{A}{N_{NS}}e^{-\beta(\Delta\epsilon_A+\epsilon_{ap})}}{1 + \frac{A}{N_{NS}}e^{-\beta(\Delta\epsilon_A)}}. \quad (1.19)$$

The examples of simple repression and simple activation show how statistical mechanical models can be applied to simple architectures. One can write quantitative models for any combination of interacting transcription factor binding sites, and such models have been written for each of the transcription factor architectures found in this work (Bintu et al., 2005). Further Refs. (Boedicker et al., 2013; Scott et al., 2010) apply the states and weights approach to the case of DNA looping in the *lac* operon.

1.4 Quantitative modeling of gene regulation

As previously mentioned a core principle of this work is the power of quantitative modeling for developing an understanding of gene regulation. As biology advances, measurements get more and more precise. Biology as a field, and gene regulation in particular has traditionally focused on qualitative questions such as the effect on phenotype of knocking out a particular protein. However, advances in measurement technology allows for quantitative models to make falsifiable predictions of a system's behavior. Additionally, analytically-derived quantitative models allow us to test our understanding of the essential mechanisms that drive a system. As an example, for regulation of transcription we typically use models that rely on the occupancy hypothesis Ackers and Johnson, 1982, namely that the probability of RNAP binding to the promoter, p_{bound} , is proportional to gene expression. The occupancy hypothesis posits that binding of RNAP or a transcription factor to a binding site

indicates that the protein is actively playing a role in transcription. This means that for RNAP, occupancy of a promoter implies that transcription is taking place; We often apply this assumption when writing models for transcriptional regulation, such as that shown for simple activation and repression above, but it is not always valid. In Phillips et al., 2012 it was found that the occupancy hypothesis could not adequately describe the mechanism of repression a single repression architecture where the repressor bound upstream of the RNAP and could bind simultaneously to the RNAP. In this case gene regulation occurred independent of the occupancy of RNAP.

1.5 Lack of current regulatory knowledge

THE REGULATORY GENOME OF *ESCHERICHIA COLI*: PROMOTER STUDIED

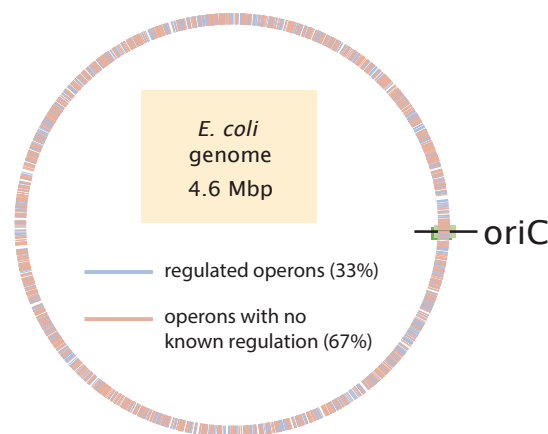


Figure 1.7: Identification of operons in *E. coli* with and without regulatory annotation. The plot identifies the genomic location of different operons with annotated TF binding sites, and those lacking regulatory descriptions. The identification of regulated operons was performed using data from RegulonDB (Gama-Castro et al., 2016), which are based on manually curated experimental and computational data. All operons listed in the database were considered, where an operon was assumed to be regulated if it had at least one transcription factor binding site associated with it.

Much of the insight we have on gene regulation relies on careful and extensive work of a few model regulatory systems (Daber and Sochor, 2011; Kuhlman et al., 2007; Buchler, Gerland, and Hwa, 2003; Vilar and Leibler, 2003); Much of the quantitative work that has come from the efforts of the Phillips group has been

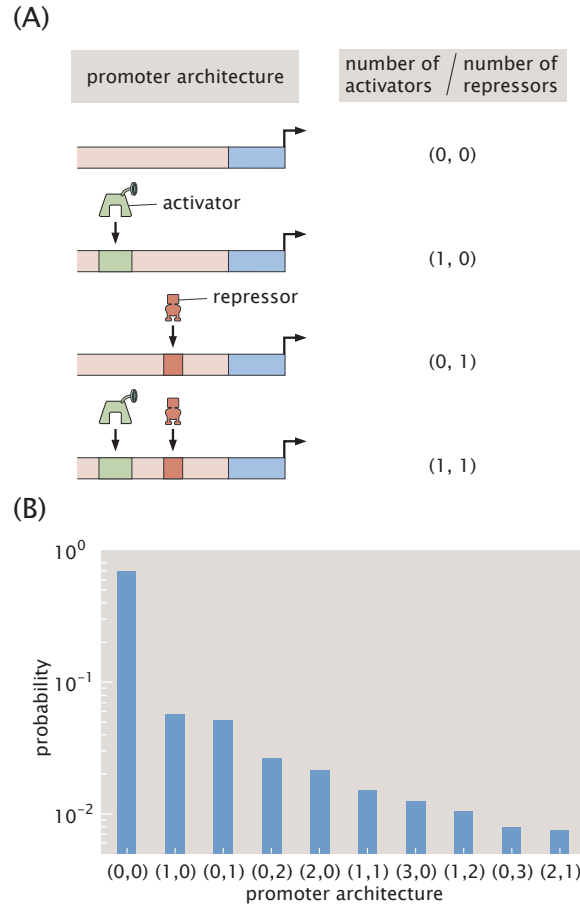


Figure 1.8: Distribution of regulatory architectures in *E. coli*. The percentage prevalence of each type of regulatory architecture in RegulonDB before the work in this dissertation. (A) Examples of some types of regulatory architectures in *E. coli*. The type of architectures are given as (A, R) where A is the number of activator binding sites and R is the number of repressor binding sites. (B) We plot the frequencies of different regulatory architectures as reported by RegulonDB (Gama-Castro et al., 2016). Note that many promoters lack complete regulatory annotations, which can often mean they are understudied rather than truly constitutive, which skews the data towards (0,0).

exclusively focused on the *lac* operon (Oehler et al., 1990; Schleif, 2010; Garcia and Phillips, 2011; Brewster, Jones, and Phillips, 2012; Boedicker et al., 2013; Brewster, Weinert, et al., 2014; Forcier et al., 2018). Other quantitative work has been focused on artificial promoters (Urtecho et al., 2019). In the case of *E. coli* and other prokaryotes, the failure to extend quantitative methods to other promoters come not from a failure of the theory of gene regulation, but rather a failure in knowledge of basic regulatory information as displayed in Fig. 1.7. While impressive advances in

molecular biology have made it possible to map thousands of gene interactions and create genetic networks for a variety of organisms. We see statistics for the different types of regulatory architectures from RegulonDB displayed in Fig. 1.8. Even so, all the knowledge of gene regulation still leave us with a regulatory landscape that is qualitative, and the vast majority of the "regulated" genes alluded to in Fig. 4.1 have none of the quantitative details, such as interaction energies between proteins, that are necessary for the formation of strong predictions about gene regulation under perturbations such as mutation or changes in growth condition. The poor state of regulatory knowledge is the primary stumbling block to applying the hard-won knowledge of quantitative models based on statistical mechanics and motivates the work in the following chapters.

Furthermore, Fig. 1.7 identifies the positions of each operon on the *E. coli* genome and whether it contains annotated transcription factor binding sites (blue) or not (red). It is striking that over half of the operons lack any listed transcription factor binding sites. One hypothesis is that the majority of operons express constitutively (i.e., no transcription factors regulate these operons). Alternatively, transcription might be controlled through changes in σ factor concentrations, which would provide an alternative mechanism of regulation. σ factors are necessary for transcription and, especially for some specialty σ factors, such as $\sigma 54$, which responds to heat shock, they can increase gene expression in response to stimuli. As another example, in stationary phase there is an increase in the cellular concentration of stationary phase sigma factor, RpoS ($\sigma 38$), which decreases the level of functional sigma factor RpoD ($\sigma 70$) and alters the genome-wide transcription output (Jishage et al., 1996). A motivation for our work and our assertion that there is a huge amount of missing regulatory knowledge is a recent proteome-wide census that was taken in *E. coli* across 22 growth conditions (Schmidt et al., 2016). In this work Schmidt *et al.* measured the copy number of more than half the *E. coli* proteome across a variety of relevant conditions such as different growth media.

As reported by Schmidt *et al.*, we also find that the GalE protein shows significantly higher expression when cells were grown in galactose, which is displayed in Fig 1.9 (A). GalE is known to be regulated, which is relieved when grown in galactose (Irani, Orosz, and Adhya, 1983; Semsey et al., 2007). We display how expression tends to vary among growth conditions. Among promoters without any known regulation, we show the expression of *dgoD* in Fig. 1.9 (B) in several carbon sources. This is only one of many examples where a protein showed a large change

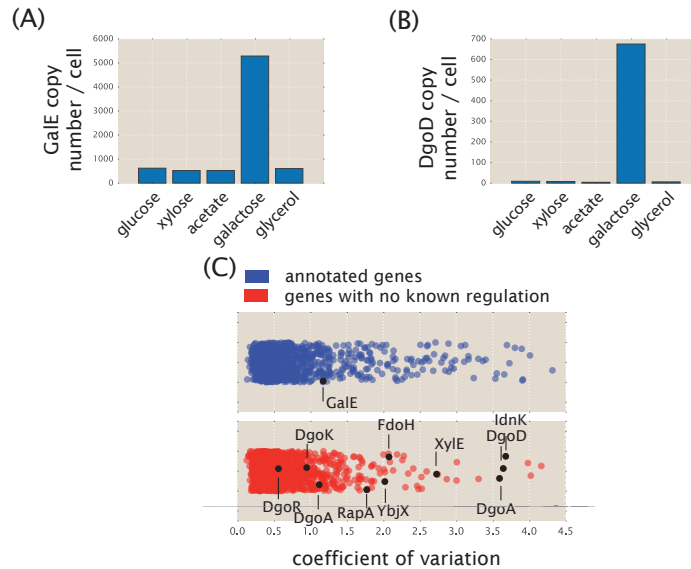


Figure 1.9: Proteome Data from Schmidt et al., 2016. (A) The expression of the *galE* gene under different growth conditions. *galE* is known to be regulated. (B) Expression of *dgoD* under different growth conditions. *dgoD* was not known to be regulated. (C) The coefficient of variation for all proteins in the Schmidt et al., 2016 dataset. Several of the genes that are discussed in detail in Chapters 3 and 4 are highlighted.

in expression level across growth conditions and suggests that there is missing regulation. In addition, we see in Fig. 1.9 (C) that the expression variability for unannotated genes appears visually almost as variable as those with known regulation, further suggesting that many of the unannotated operons have missing regulation. In Chapter 3, we prove there was missing regulation for the *xylE* and *dgoRKADT* operons that were displayed in 1.9 (C). Additionally, in Chapter 4 we show that this missing regulation is widespread across the more than 100 *E. coli* promoters that were studied.

BIBLIOGRAPHY

- Ackers, Gary K and Alexander D Johnson (1982). “Quantitative model for gene regulation by A phage repressor”. en. In: *Proc. Natl. Acad. Sci. USA*, p. 5.
- Avery, Oswald, Colin M. MacLeod, and Maclyn McCarty (Feb. 1944). “studies on the chemical nature of the substance inducing transformation of pneumococcal types”. In: *The Journal of Experimental Medicine* 79.2, pp. 137–158.
- Bintu, Lacramioara et al. (Apr. 2005). “Transcriptional regulation by the numbers: models”. en. In: *Current Opinion in Genetics & Development*. Chromosomes and expression mechanisms 15.2, pp. 116–124. doi: 10.1016/j.gde.2005.02.007.
- Boedicker, James Q et al. (Nov. 2013). “DNA sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation”. en. In: *Physical Biology* 10.6, p. 066005. doi: 10.1088/1478-3975/10/6/066005.
- Bonocora, Richard P. and Joseph T. Wade (2015). “ChIP-Seq for Genome-Scale Analysis of Bacterial DNA-Binding Proteins”. en. In: *Bacterial Transcriptional Control: Methods and Protocols*. Ed. by Irina Artsimovitch and Thomas J. Santangelo. Methods in Molecular Biology. New York, NY: Springer, pp. 327–340. doi: 10.1007/978-1-4939-2392-2_20.
- Brewster, Robert C., Daniel L. Jones, and Phillips (Dec. 2012). “Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*”. en. In: *PLoS Computational Biology* 8.12. Ed. by Erik van Nimwegen, e1002811. doi: 10.1371/journal.pcbi.1002811.
- Brewster, Robert C., Franz M. Weinert, et al. (Mar. 2014). “The Transcription Factor Titration Effect Dictates Level of Gene Expression”. en. In: *Cell* 156.6, pp. 1312–1323. doi: 10.1016/j.cell.2014.02.022.
- Buchler, Nicolas E., Ulrich Gerland, and Terence Hwa (Apr. 2003). “On schemes of combinatorial transcription logic”. en. In: *Proceedings of the National Academy of Sciences* 100.9, pp. 5136–5141. doi: 10.1073/pnas.0930314100.
- Cisse (2020). “Coactivator condensation at super-enhancers links phase separation and gene control | Science”. In: ().
- Cisse et al. (Aug. 2013). “Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells”. en. In: *Science* 341.6146, pp. 664–667. doi: 10.1126/science.1239053.
- Compan, Inès and Danlèle Touati (1994). “Anaerobic activation of *arcA* transcription in *Escherichia coli*: roles of Fnr and ArcA”. en. In: *Molecular Microbiology* 11.5, pp. 955–964. doi: 10.1111/j.1365-2958.1994.tb00374.x.
- Crick, Francis (1961). “General Nature of the Genetic Code for Proteins”. en. In: p. 3.

- Daber, Robert and Mitchell Sochor Matthew and Lewis (May 2011). “Thermodynamic Analysis of Mutant lac Repressors”. en. In: *Journal of Molecular Biology*. The Operon Model and its Impact on Modern Molecular Biology 409.1, pp. 76–87. DOI: 10.1016/j.jmb.2011.03.057.
- Easton, A M and S R Kushner (Dec. 1983). “Transcription of the uvrD gene of *Escherichia coli* is controlled by the lexA repressor and by attenuation.” In: *Nucleic Acids Research* 11.24, pp. 8625–8640. DOI: 10.1093/nar/11.24.8625.
- Forcier, Talitha L et al. (2018). “Measuring cis-regulatory energetics in living cells using allelic manifolds”. en. In: p. 28.
- Gama-Castro, Socorro et al. (Jan. 2016). “RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond”. en. In: *Nucleic Acids Research* 44.D1, pp. D133–D143. DOI: 10.1093/nar/gkv1156.
- Garcia and Phillips (July 2011). “Quantitative dissection of the simple repression input-output function”. en. In: *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–12178. DOI: 10.1073/pnas.1015616108.
- Ingolia et al. (Apr. 2009). “Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling”. en. In: *Science* 324.5924, pp. 218–223. DOI: 10.1126/science.1168978.
- Irani, Meher H., Laszlo Orosz, and Sankar Adhya (Mar. 1983). “A control element within a structural gene: The gal operon of *Escherichia coli*”. en. In: *Cell* 32.3, pp. 783–788. DOI: 10.1016/0092-8674(83)90064-8.
- Jishage, M et al. (1996). “Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions.” en. In: *Journal of bacteriology* 178.18, pp. 5447–5451. DOI: 10.1128/JB.178.18.5447-5451.1996.
- Kinney and Callan (May 2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. en. In: *Proceedings of the National Academy of Sciences* 107.20, pp. 9158–9163. DOI: 10.1073/pnas.1004290107.
- Kinney and McCandlish (Aug. 2019). “Massively Parallel Assays and Quantitative Sequence–Function Relationships”. en. In: *Annual Review of Genomics and Human Genetics* 20.1, pp. 99–127. DOI: 10.1146/annurev-genom-083118-014845.
- Kuhlman et al. (Apr. 2007). “Combinatorial transcriptional control of the lactose operon of *Escherichia coli*”. en. In: *Proceedings of the National Academy of Sciences* 104.14, pp. 6043–6048. DOI: 10.1073/pnas.0606717104.
- Kumar, Rahul and Kazuyuki Shimizu (2011). “Transcriptional regulation of main metabolic pathways of cyoA, cydB, fnr, and fur gene knockout *Escherichia coli* in C-limited and N-limited aerobic continuous cultures”. en. In: *Microbial Cell Factories* 10.1, p. 3. DOI: 10.1186/1475-2859-10-3.

- Melnikov, Alexandre et al. (Mar. 2012). “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay”. en. In: *Nature Biotechnology* 30.3, pp. 271–277. DOI: 10.1038/nbt.2137.
- Oehler et al. (Apr. 1990). “The three operators of the lac operon cooperate in repression.” en. In: *The EMBO Journal* 9.4, pp. 973–979. DOI: 10.1002/j.1460-2075.1990.tb08199.x.
- Pachter (Aug. 2013). *Seq. en.
- Phillips (2015). “Napoleon Is in Equilibrium”. In: *Annual Review of Condensed Matter Physics* 6.1, pp. 85–111. DOI: 10.1146/annurev-conmatphys-031214-014558.
- Phillips et al. (Oct. 2012). “Organization of Biological Networks”. en. In: *Physical Biology of the Cell*. DOI: 10.1201/9781134111589-19.
- Schleif, Robert (Sept. 2010). “AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action”. en. In: *FEMS Microbiology Reviews* 34.5, pp. 779–796. DOI: 10.1111/j.1574-6976.2010.00226.x.
- Schmidt, Alexander et al. (Jan. 2016). “The quantitative and condition-dependent *Escherichia coli* proteome”. en. In: *Nature Biotechnology* 34.1, pp. 104–110. DOI: 10.1038/nbt.3418.
- Scott, M. et al. (Nov. 2010). “Interdependence of Cell Growth and Gene Expression: Origins and Consequences”. en. In: *Science* 330.6007, pp. 1099–1102. DOI: 10.1126/science.1192588.
- Semsey, Szabolcs et al. (2007). “Signal integration in the galactose network of *Escherichia coli*”. en. In: *Molecular Microbiology* 65.2, pp. 465–476. DOI: 10.1111/j.1365-2958.2007.05798.x.
- Urtecho, G. et al. (2019). “Systematic Dissection of Sequence Elements Controlling sigma70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*”. In: *Biochemistry* 58.11, pp. 1539–1551. DOI: 10.1021/acs.biochem.7b01069.
- Vilar, José M. G. and Stanislas Leibler (Aug. 2003). “DNA Looping and Physical Constraints on Transcription Regulation”. en. In: *Journal of Molecular Biology* 331.5, pp. 981–989. DOI: 10.1016/S0022-2836(03)00764-2.
- Watson, J. D. and Crick (Apr. 1953). “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid”. en. In: *Nature* 171.4356, pp. 737–738. DOI: 10.1038/171737a0.
- Yona, Avihu H., Eric J. Alm, and Jeff Gore (Apr. 2018). “Random sequences rapidly evolve into de novo promoters”. en. In: *Nature Communications* 9.1, pp. 1–10. DOI: 10.1038/s41467-018-04026-w.

Chapter 2

MAPPING DNA SEQUENCE TO TRANSCRIPTION FACTOR BINDING ENERGY *IN VIVO*.

A version of this chapter originally appeared as S. L. Barnes, N. M. Belliveau, W. T. Ireland, M. J. Sweredoski, J. B. Kinney, and R. Phillips (2018). Mapping DNA sequence to transcription factor binding energy in vivo. PLOS Computational Biology, <http://doi.org/10.1371/journal.pcbi.1006226>.

Author contribution note: for this chapter, I (WI) assisted with experimental design, data analysis, and manuscript writing.

2.1 Introduction

High-throughput sequencing has delivered on the promise that we can sequence the genome of nearly any species at will. The amount of genome data available is already enormous and will only continue to grow. However, this mass of data is nearly useless without the appropriate methods of analyzing it. Despite decades of research, genomic data still defies our efforts to “read” it. When faced with an entirely new genome, we can guess that a stretch of DNA contains a gene, and then use the codon table for amino acids to translate that hypothetical gene into an amino acid sequence. In some cases, we can also guess at, or measure with techniques like RACE (Mendoza-Vargas et al., 2009), the locations of transcription start sites. In some cases we can even make guesses as to the locations of transcription factor binding sites, but these guesses tell us little about how the details of a putative site lead to its downstream effects on gene expression. A more detailed understanding of the sequence dependence of gene expression and transcription factor binding sites is needed in order to improve the accuracy of such predictions. An important avenue for developing this level of understanding is to propose models that map sequence to function and perform experiments that test these models, which will hopefully lead to an understanding of gene regulation.

Past efforts have found it difficult to unravel the mysteries of gene regulation, even on a single gene, much less the thousands of genes that it would take to gain a full regulatory understanding of even *E. coli*. Over half of the genes in *E. coli*, which is arguably the best-understood model organism, lack any regulatory annotation

(Gama-Castro et al., 2016). Those operons whose regulation is well described such as the *lac*, *rel*, and *mar* operons (Oehler et al., 1990; Grainger et al., 2005; Alekshun and Levy, 1997) required decades of work, often involving laborious genetic and biochemical experiments (Minchin and Busby, 2009). A wide variety of new techniques have been proposed and implemented to simplify the process of determining how a gene is regulated. ChIP-based methods such as ChIP-chip and ChIP-seq make it possible to determine the genome-wide binding locations of individual transcription factors of interest. Massively parallel reporter assays (MPRAs) have made it possible to read out transcription factor binding position and occupancy *in vivo* with base-pair resolution, and provide a means for analyzing non-binding features such as “insulator” sequences (Levo, Avnit-Sagi, et al., 2017; Melnikov et al., 2012; Levy et al., 2017). *in vitro* methods such as protein-binding microarrays (Berger et al., 2006), SELEX (Fields et al., 1997; Jolma et al., 2013), MITOMI (Maerkl, 2007; Shultzaberger et al., 2012), and binding assays performed in high-throughput sequencing flow cells (Jung et al., 2017; Nutiu et al., 2011) have made it possible to measure transcription factor affinity to a broad array of possible binding sites and develop detailed records of transcription factor sequence specificities.

In spite of this progress, it remains difficult to integrate the various aspects of transcriptional regulation revealed by such experiments into a cohesive understanding of a given promoter or transcription factor. While *in vitro* methods may provide accurate measurements of transcription factor sequence specificities and binding affinities, including insight into the effects of flanking sequences (Dror et al., 2015; Levo, Zalekvar, et al., 2015), they cannot fully account for the *in vivo* consequences of binding site context and interactions with other proteins. Current *in vivo* methods for determining transcription factor binding affinities, such as bacterial one-hybrid (Christensen et al., 2011; Xu and Noyes, 2015), require a restructuring of the promoter so that it no longer resembles its genomic counterpart. Additionally, while computational efforts to “read” the genome by scanning for DNA sequences that resemble known transcription factor binding sites provide a promising avenue for understanding transcriptional regulation in its native context, these efforts frequently produce false positives (Weirauch et al., 2013; Djordjevic, 2003) as we also see during Chapter 4. Furthermore, a common assumption underlying many of these methods is that transcription factor occupancy in the vicinity of a promoter implies regulation, but it has been shown that occupancy cannot accurately predict the effect of a transcription factor on gene regulation (Garcia, Sanchez, et al., 2012;

Wunderlich and Mirny, 2009).

An ideal technique would be capable of interrogating multiple aspects of transcriptional regulation at once, from locating transcription factor binding sites to identifying the sequence specificity of these binding sites. As previously noted, massively parallel reporter assays have shown a great deal of promise for this reason. In Brewster, Jones, and Phillips, 2012, we showed that the MPRA Sort-Seq (Kinney and Callan, 2010), combined with a simple linear model for protein-DNA binding specificity, can be used to accurately predict the binding energies of multiple RNAP binding site mutants, serving as a jumping off point for the use of such models as a quantitative tool in synthetic biology. Here we adopt a similar philosophy to explore whether this technique can be more broadly applied to other regulatory components such as transcription factor binding sites.

Specifically, we use Sort-Seq to map sequence to binding energy for the repressor-binding site interaction, and we rigorously characterize the variables that must be considered in order to obtain an accurate sequence-binding energy map. We show how such a mapping can be used to characterize how sequence controls protein binding and, ultimately, gene expression. We validate the approach via comparisons with microscopy data and explore the limits of the simple linear models of binding energy that we use. As concrete applications of this approach, we show that our sequence-energy mapping can be used to precisely design a series of binding sites with a hierarchy of precisely controlled binding energies. With this suite of different binding energies in hand, we then show how those binding sites can be used to design a wide range of induction responses with different phenotypic properties such as leakiness, dynamic range and $[EC_{50}]$. Finally, we use Sort-Seq when single amino acid perturbations to the LacI protein have been introduced, and we characterize how this affects the mapping of DNA sequence specificity. This broad collection of case studies provides a rigorous test of the quantitative mapping between regulatory sequence and function offered by the Sort-Seq approach.

2.2 Results

In order to map regulatory sequence to binding energy *in vivo*, we applied Sort-Seq (Kinney and Callan, 2010) to synthetically constructed promoters with binding sites for RNA polymerase (RNAP) and *lac* repressor (LacI). As shown in Fig. 2.1(A), Sort-Seq works by first generating a library of cells, each of which contains a mutated promoter that drives expression of GFP from a low copy plasmid (5-10 copies per

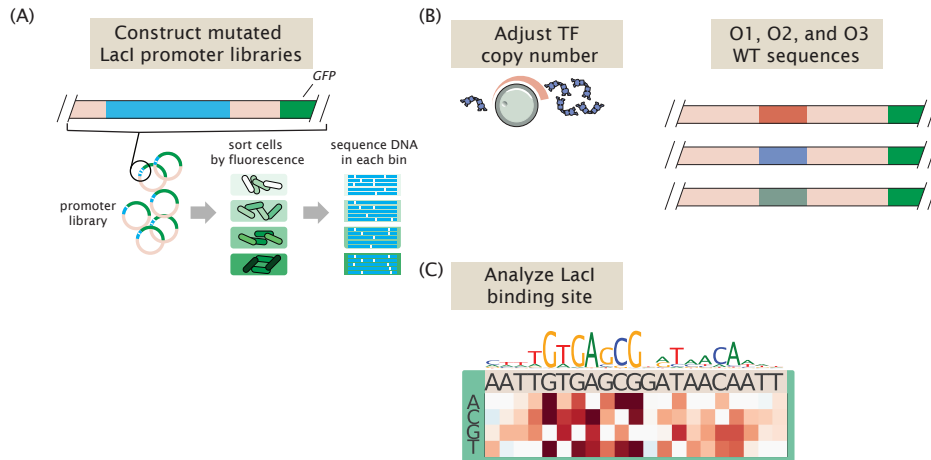


Figure 2.1: Process flow for using Sort-Seq to obtain energy matrices. (A) A simple repression motif was designed in which a LacI repressor binding site is placed immediately downstream of the RNAP site. RNAP binding probability will be proportional to GFP production. The RNAP and LacI binding sites were both randomly mutated at a rate of approximately 10% and the resulting plasmid library was transformed into cells such that each cell contains a different mutant. We then sort the cell population into bins based on fluorescence level, and then sequence the cells in each bin to map sequence to expression. (B) We analyze simple repression constructs using each of the three lac operators that are found in *E. coli*, O1, O2, and O3, and performed Sort-Seq in *E. coli* strains with mean copy numbers of LacI per cell of 22 ± 4 , 60 ± 20 , 124 ± 30 , 260 ± 40 , 1220 ± 160 , and 1740 ± 340 (using strains from Garcia and Phillips, 2011). The resulting Sort-Seq data was used to infer energy matrices that describe the sequence-dependent repression by LacI. An example energy matrix and sequence logo (G. D. Stormo, 2000) are shown for LacI, with the convention that the wild-type nucleotides have zero energy.

cell; Lutz, 1997). GFP is a fluorescent protein that allows expression level of the protein to be observed by measuring the fluorescence level. We use fluorescence-activated cell sorting (FACS) to sort that library of cells into multiple bins gated by their fluorescence level and then sequence the mutated plasmids from each bin. Binding by LacI to the promoter physically occludes binding by RNAP (Ackers and Johnson, 1982; Buchler, Gerland, and Hwa, 2003), and mutations to both binding site sequences will influence what bin each cells is sorted into.

One of the important aspects demonstrated by Kinney and Callan, 2010, is that we can use the large sequence data set from Sort-Seq (0.5-2 million sequences) to perform information-based modeling and extract quantitative information from the data. In particular, it is possible to infer energy matrix models that describe the sequence dependent energy of interaction between transcription factors and their

binding sites (Kinney and Callan, 2010; Ireland and Kinney, 2016). Here we set out to test the accuracy of the models that come from Sort-Seq experiments in the context of the simple repression architecture (Bintu et al., 2005), with repression by LacI as noted above.

In order to be more representative of the range in both transcription factor and protein-DNA binding energies observed in *E. coli* more generally, but also to test the capabilities of the approach more broadly, we constructed a set of strains with a range of repressor copy numbers and DNA binding energies. Both of these factors are key determinants of gene expression for a simple repression architecture as can be seen in Eq. 1.15. We performed a set of separate Sort-Seq experiments in *E. coli* with mean LacI dimer copy numbers ranging from 22-1740 copies per cell (Fig. 2.1(B)). We varied the binding site sequence of the LacI binding site in our promoter library, using the three natural sites found at the *lac* operon (O1 with binding energy, $-15.3 k_B T$; O2, the second strongest, $-13.9 k_B T$; and O3 the weakest at $-9.7 k_B T$ (Garcia and Phillips, 2011)).

Sequence-dependent thermodynamic model of the simple repression architecture

We begin by defining the thermodynamic model of simple repression that we will apply to our Sort-Seq data. We will also define energy matrices that describe the sequence-dependent interaction energies of RNAP and LacI to their binding sites.

We consider a cell with P copies of RNAP per cell and R copies of LacI per cell, and begin by enumerating all possible states of the promoter and their corresponding statistical weights. As shown in Fig. 2.2, the promoter can either be empty, occupied by RNAP, or occupied by LacI. In addition to these specific binding sites, we assume that there are $N_{NS} = 4.6 \times 10^6$ non-specific binding sites elsewhere on the chromosome where RNAP and LacI may bind non-specifically. We define our reference energy such that all specific binding energies are measured relative to the average non-specific binding energy. For simplicity, our model explicitly ignores the complexity of the distribution of non-specific binding affinities in the genome and makes the assumption that a single parameter can capture the energy difference between our binding site of interest and the average site in the reservoir.

Thermodynamic models of transcription assume that gene expression is proportional to the probability that the RNAP is bound to the promoter p_{bound} , and as we have found in Chapters 1, this is given by

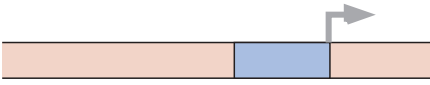

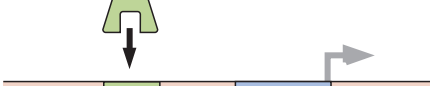

State	Weight
	1
	$\frac{P}{N_{NS}} e^{-\beta \Delta \epsilon_P}$
	$\frac{A}{N_{NS}} e^{-\beta \Delta \epsilon_A}$
	$\frac{PA}{N_{NS}} e^{-\beta (\Delta \epsilon_A + \Delta \epsilon_P + \epsilon_{AP})}$

Figure 2.2: States and weights for the simple repression motif. There are P RNA polymerase (blue) and R repressors (red) per cell that compete for binding to a promoter of interest. The difference in energy between a repressor bound to the promoter of interest versus another non-specific positions elsewhere on the DNA equals $\Delta \epsilon_R$; the P RNA polymerase have a corresponding energy difference $\Delta \epsilon_P$ relative to non-specific binding on the DNA. N_{NS} represents the number of non-specific binding sites for both RNA polymerase and repressor.

$$p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \epsilon_P}}{1 + \frac{p_A(c)R}{N_{NS}} e^{-\beta \Delta \epsilon_R} + \frac{P}{N_{NS}} e^{-\beta \Delta \epsilon_P}} \quad (2.1)$$

with $\beta = \frac{1}{k_b T}$, where k_b is the Boltzmann constant and T is the temperature of the system. Here we have included the allosteric aspect of LacI through the term, $p_A(c)$, which indicates the fraction of active LacI in the presence of inducer. c denotes the concentration of inducer present in the cell ($p_A(c) = 1$ when no inducer is present).

We describe the sequence-dependent binding energies for RNAP, $\Delta \epsilon_P$, and LacI, $\Delta \epsilon_R$, using linear energy matrix models. We define the binding energy associated with each protein z , $\Delta \epsilon_z$ ($z = P$ for RNAP, and $z = R$ for LacI), by

$$\Delta \varepsilon_z = \alpha \varepsilon_{z,mat} + \Delta \varepsilon_{z,wt}, \quad (2.2)$$

where $\varepsilon_{z,mat}$ is the energy value obtained by summing the matrix elements associated with a sequence (further defined below), α_z is a scaling factor that converts the matrix values into $k_b T$ units, and $\Delta \varepsilon_{z,wt}$ is the binding energy associated with the wild-type operator.

Energy matrices treat each base pair position j along a binding site as contributing a certain amount to the binding energy. Mathematically the energy matrix is described by a $4 \times L$ matrix, where each column j of matrix parameters will represent the energies for each nucleotide $i = A, C, G, T$ associated with position j of the binding site. For example, index $(i=C, j=3)$ represents the energy parameter for nucleotide C at position 3. The binding energy of a sequence from an energy matrix will then be given by

$$\varepsilon_{z,mat} = \sum_{i=1}^L \sum_{j=A}^T \theta_{ij} \cdot s_{ij}, \quad (2.3)$$

where θ_{ij} represents the parameters of the energy matrix and s_{ij} , at position j of the binding site, with base identity i , and the subscript z represents the matrix either being a LacI or RNAP matrix. Although we only refer to linear matrices in 2.3, these models can be extended to allow for non-additive contributions from each position, though linear models appear to be sufficient to describe transcription factor binding in bacteria in general (Berg and Hippel, 1987; Benos, Bulyk, and Gary D. Stormo, 2002; Brewster, Jones, and Phillips, 2012). By convention, we have fixed the values of the matrix positions associated with the wild-type sequence to 0 $k_b T$, so that $\varepsilon_{mat} = 0$ for a wild-type sequence. Thus, $\alpha \varepsilon_{mat}$ can be interpreted as the change in binding energy relative to the wild-type energy caused by specific mutations in the sequence of interest.

Inferring models of the simple repression architecture using Sort-Seq

We use the MPATHic software to infer the parameters of the energy matrices and thermodynamic parameters of p_{bound} (Kinney and Callan, 2010; Ireland and Kinney, 2016). The software uses Markov-Chain Monte Carlo (MCMC) to determine the set of parameters that maximize the mutual information between the distribution of sequences in the binned sequence data and the model's predictions. More specifically, the inference approach samples the probability distribution

$$p(\theta|S, b) \propto 2^{NI(b, \text{model predictions})}. \quad (2.4)$$

Here θ is the set of model parameters that define our model (e.g. entries in the energy matrices), S, b represents our data set of sequences S and the sorted bin b where they were found. N is the number of sequences in the data, and $I(b, \text{model predictions})$ is the mutual information between the distribution of binned sequences and the model's predictions, which we discuss further below.

Due to the computational burden of fitting a large number of parameters by MCMC (all parameters), we find it convenient to first infer the energy matrices (in arbitrary units) for LacI and RNAP from the Sort-Seq data. Fig 2.3 (A) summarizes the result for one of the LacI energy matrices (using an O1 binding site library, and *E. coli* strain with $R = 1740$ LacI per cell). Fig. 2.3 (B) shows an energy matrix after the energy scale is fixed (in k_bT units) using the thermodynamic model in 2.1 to make model predictions. Mutual information is estimated from the joint probability distribution between model prediction and binned sequence data, which is estimated by performing kernel density estimation. Note that in this instance, we are estimating a joint distribution to calculate the mutual information between sequence bin and energy prediction, $I(b, \text{energy (a.u.)})$. We repeat this procedure to generate an energy matrix for the RNAP binding site.

With our energy matrices in hand, we use Sort-Seq sequence data to determine the scaling parameter of Eq. 2.2 by fitting the data against the thermodynamic model defined by Eq. 2.1. In fitting the thermodynamic model we must use a parallel tempering Monte Carlo method. This is because the likelihood landscape from fitting the scaling parameters can be quite rough, with many local maxima. Parallel tempering uses multiple MCMC runs with different "temperatures". In other words, there will be some MCMC chains that are very permissive in which MCMC steps they accept and so widely explore the parameter landscape. These high temperature walkers are able to escape local minima in the likelihood landscape. The low temperature walkers are able to locally explore areas of high likelihood well. Parallel tempering algorithms periodically exchange the model parameters between the MCMC chains running at different temperatures with a probability related to the relative difference in temperature. We use the Emcee package to perform the parallel tempering algorithm (Foreman-Mackey et al., 2013).

In Fig. 2.4 (A) we summarize the energy matrices for LacI for the strains with

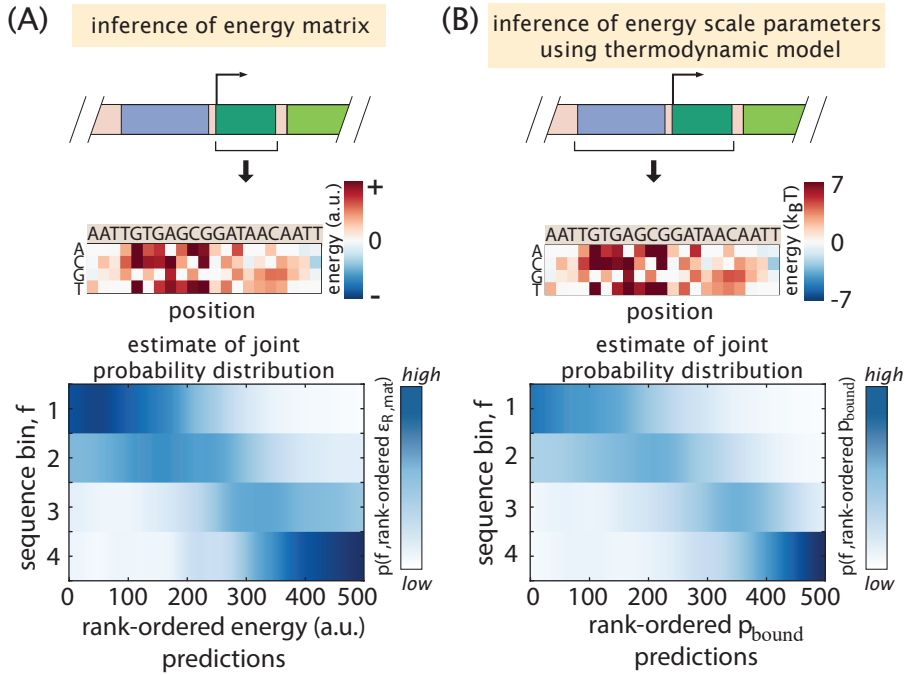


Figure 2.3: Inference of LacI energy matrices. (A) Using the aligned sequence data for the LacI binding site, information-based modeling was performed with the MPAthic software (Ireland and Kinney, 2016) to determine the parameters of the LacI energy matrix (in arbitrary units). By convention, the energies are defined such that the O1 wild-type sequence has zero energy. Kernel density estimation was performed to estimate the joint probability distribution between sequence bin f and rank-ordered energy predictions from the inferred matrix. (B) Sort-Seq data was fit to the thermodynamic Eq. 2.1, where binding energies were calculated from the separately inferred energy matrices for LacI and RNAP. The entire promoter sequence from each mutated sequence was used in this inference. This allowed determination of the scaling factors for binding by LacI and the energy matrix shown in absolute $k_B T$ energy units. A joint probability distribution between sequence bin f and rank-ordered predictions of p_{bound} is shown using the inferred model. Data is from the Sort-Seq experiment using an O1 LacI binding site and performed in a strain with $R = 1740$ repressor copies per cell.

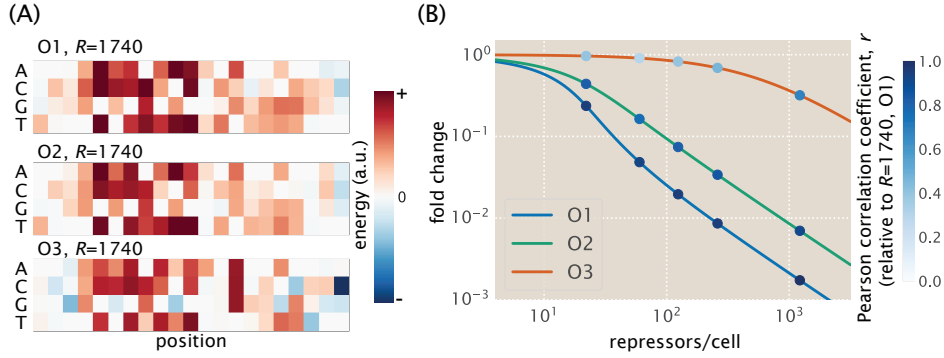


Figure 2.4: Energy matrices for the natural lac operators from Sort-Seq data. (A) Energy matrix models are shown for the LacI binding site from experiments performed with O1, O2, and O3 libraries, and in strains with $R = 1740$ repressor copies per cell. All energy matrices are plotted such that an O1 binding site sequence will have zero energy. (B) Pearson correlation coefficients were calculated relative to the energy matrix found using the O1 library in a strain with $R = 1740$ repressor copies per cell. Each marker represents the correlation coefficient for a matrix from a separate Sort-Seq experiment. Data is overlaid on a plot of expected expression fold-change (calculated assuming 10 plasmid copies per cell (Weinert et al., 2014)) to provide a reference for the expected influence of LacI on expression under each particular Sort-Seq experiment.

the highest repressor copy number, $R = 1740$. Here we plot the energy matrices generated from each operator and compare the sequence specificity of each matrix. We find that the energy matrices from the O1 and O2 binding site data are quite similar, while the matrix from the O3 binding site data is somewhat less consistent (Pearson correlation coefficients: $r = 0.91$ between O1 and O2; $r = 0.69$ between O1 and O3).

The entire set of LacI matrices generated from the Sort-Seq experiments are summarized in Fig. 2.4 (B). Here we calculate the correlation of each matrix (relative to the $R = 1740$, O1 energy matrix), and overlay these values on a plot of the expected fold-change as a function of repressor copy number. Fold-change here refers to the ratio of gene expression in the presence of repressor relative to expression in the absence of repressor and provides a useful measure for the extent of repression expected by LacI in each Sort-Seq experiment. We find each matrix from the O1 and O2 binding site data sets to be quite consistent. Notably however, those from the O3 binding site data sets are less similar. Given the low repression expected by LacI in strains with an O3 binding site, this result may be due to the Sort-Seq data containing less information content associated with binding of LacI. Though it is

also useful to note that we also find some correlation among matrices based on the same binding library ($r > 0.94$ across O1 matrices; $r > 0.91$ across O2 matrices, and $r > 0.80$ across O3 matrices).

Sort-Seq energy matrices provide accurate prediction of LacI binding energy

In order to test the binding energy predictions that are provided by our LacI energy matrices, we constructed a set of simple repression constructs where the O1 binding site was mutated at 1, 2, or 3 positions (summarized in Table 2.1). These were placed into our *E. coli* strains containing different LacI copy numbers ($R = 22 \pm 4$, 60 ± 20 , 124 ± 30 , 260 ± 40 , 1220 ± 160 , and 1740 ± 340 , where errors denote standard deviation of at least three replicates as measured in (Garcia and Phillips, 2011), and measured expression as a function of transcription factor concentration for each of the designed LacI binding sites.

Here we find it more convenient to use the fold-change in gene expression instead of expression alone. As we noted earlier fold-change is defined as the ratio of gene expression in the presence of repressor relative to expression in the absence of repressor (i.e. constitutive expression), namely

$$\text{fold-change} = \frac{p_{\text{bound}}(R > 0)}{p_{\text{bound}}(R = 0)}, \quad (2.5)$$

where p_{bound} was defined in Eq. 2.1. In section 1.3 we derived that, under the weak promoter approximation, this reduces to the form

$$\text{fold-change} \approx \left(1 + p_A(c) \frac{R}{N_{NS}} e^{-\beta \Delta \epsilon_R} \right)^{-1}. \quad (2.6)$$

For now we are only concerned with the case where no inducer is present in the growth media (i.e. where $p_A(c) = 1$). Using our LacI energy matrix to predict $\Delta \epsilon_R$, we find that we can make parameter-free predictions of fold-change for each LacI binding site sequence as a function of the repressor copy number associated with each of our *E. coli* strains.

We use flow cytometry to measure fluorescence of each strain, as explained more thoroughly in 2.4. Briefly, cells were grown to exponential phase in M9 minimal media with 0.5% glucose. Following a 1:10 dilution in fresh media, the fluorescence was measured by flow cytometry and automatically gated to include only single-cell measurements. We then calculated fold-change from the mean fluorescence level of

LacI binding site sequence	prediction (k_bT)
AATTGTGAGCGG A ACAATT	-11.9
AATTGTGAGCG C ATAACAATT	-15.6
AATTGTGAGCGGAT C ACAATT	-15.2
AATTGTGAGCGG A AACAATT	-11.5
AATTG C GAGCGGATAACAATT	-10.0
AATTGTGAG G GGATAACAATT	-12.2
AATTGTGAGCGGAT T CAATT	-12.8
AATTGTGAG C AGATAACAATT	-9.8
AATTGTGAGAGGATAACAATT	-6.3
AAATGTGAGCGGG T AACAATT	-14.6
AATTGTGAGCGGG T AACA A CT	-13.6
AAATGTGAGCGGATAACA A CT	-13.3
AATTGTGAGCG A GTAACAATT	-14.0
A TTTGTGAGCGG A ACAATT	-11.9
C ATTGTGAGCG C ATAACAATT	-15.3
AATTGTGAGCGG A CACAATT	-11.7
AATTGTGAGCGG A TACAATT	-9.6
AATTG C GAGCGGATAACAA T	-10.5
AATTGTGAG G GGATAACA T C	-14.1
AA A TGTGAGCG A GTAACAATT	-13.6
AATTGTGAGCG A ATAACA A CC	-14.6
AAATGTGAGCG A ATAACA A CT	-12.2
AATTGTGAGCG A GTAACA A CT	-12.6
A TTTGTGAGCG A AGACAATT	-10.8
AATTGTGAGCGG A CACAAT G	-12.3
AATTGTGAGCGGG A TACAATT	-9.5
AATTGT C AGCGGATAACAA A G	-11.2
AATTGTGAGGG T ATAACAATC	-13.5

Table 2.1: Summary of LacI binding site mutant energy prediction for designed O1 sites.

The binding energies displayed are the average of inferred matrices. Mutated bases are in bold.

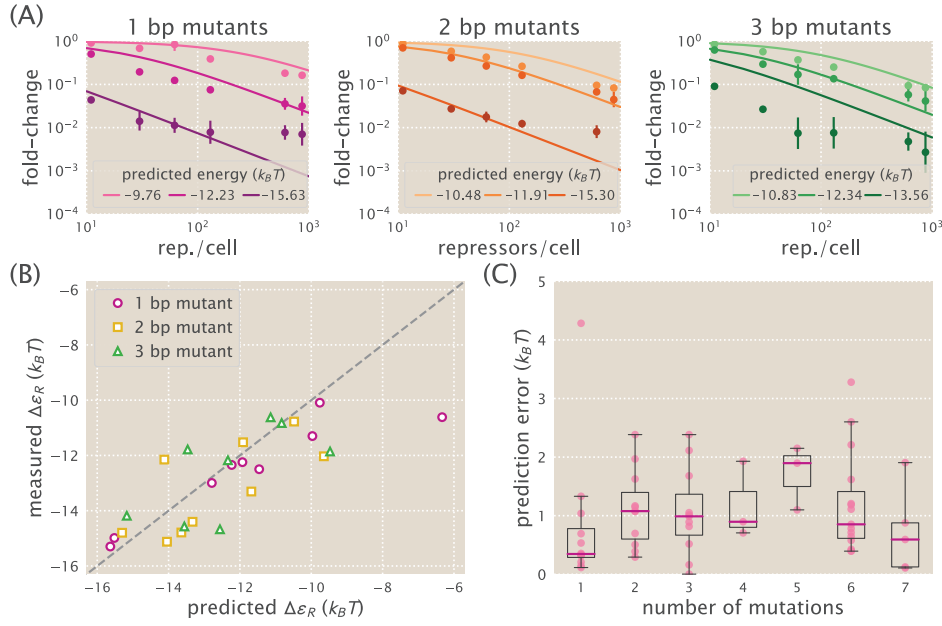


Figure 2.5: Fold-change data reflects expected values from predicted fold change curves. (A) Fold-change data were obtained for each of the mutant operators by measuring their respective fluorescence levels at multiple LacI copy numbers. The solid lines in each plot represent the expected fold-change curve for each binding energy as predicted by the O1 energy matrix. A subset of data sets are shown for the 1 bp (left), 2 bp (middle), and 3 bp (right) mutants. Approximately 30 mutants were measured in total, with five replicate measurements performed for each strain. Predicted energies are based on the average predictions from the different O1 energy matrices. (B) The measured binding energy values $\Delta\epsilon_R$ (y axis) are plotted against binding energy values predicted from an energy matrix derived from the O1 operator (x axis). While the quality of the binding energy predictions does appear to degrade as the number of mutations relative to O1 is increased, the O1 energy matrix is still able to approximately predict the measured values. (C) Binding energies for each mutant were predicted using both the O1 and O2 energy matrices and compared against measured binding energy values. The amount of error associated with each of these predictions is plotted here against the number of mutations relative to the wild-type sequence whose energy matrix was used to make the prediction. For sequences with 4 or fewer mutations, the median prediction error is consistently lower than $1.5 k_B T$.

each strain relative to a strain where LacI has been deleted. In Fig 2.5 (A) we show fold-change measurements for a subset of the 1 bp, 2 bp, and 3 bp mutants, overlaid with the parameter-free curves using our LacI energy matrix predictions of $\Delta\epsilon_R$.

Since we performed fold-change measurements for each O1 mutant at several repressor copy numbers, it was also possible to use these measurements to directly

estimate the LacI binding energies for each binding site sequence. In Fig 2.5 (B) we compare the measured binding energies against those predicted by our LacI energy matrix. For single base pair mutations most predictions are accurate to within $1 k_bT$, with many predictions differing from the measured values by less than $0.5 k_bT$. Though we do note that one of the sequences whose predicted binding was $-6.3 k_bT$, was instead found to have a binding energy of about $-10.5 k_bT$. Predictions are less accurate for 2 bp or 3 bp mutations, although the majority of these predictions are still within $1.5 k_bT$ of the measured value.

While not completely unexpected, we find that the quality of matrix predictions decreased as we predict the energy of sequences further from the wild type sequence of the binding site used to generate the energy matrix. To evaluate predictions for a wide variety of sequences, we made predictions using energy matrices made from both the O1 and O2 wild type operator sequences. The wild type O2 matrix has 5 mutations relative to O1. As a result, the tested sequences measure vary by many mutations relative to the wild type. As shown in Fig 2.5 (C), we find that predictions remain relatively accurate for mutants that have as many as 4 differences from the wild type sequence, with mean deviation of $1.5 k_bT$ or less. For a system with $R = 60$ LacI dimers, this mismatch in binding energy would imply that a prediction of fold-change would be inaccurate by $\approx 0.10 - 0.35$ (depending on the mutant binding site). by contrast, the median mismatch of $0.5 k_bT$ shown for 1 bp mutants implies that our fold-change predictions are only inaccurate by $0.04 - 0.12$, highlighting that predicted binding energies for single-point mutations will be far more reliable.

Regulatory sequence can be used to tune the simple repression induction curve.

A common desire in synthetic biology is to design regulatory circuits that provide specific input-output characteristics. A common strategy to design output levels is to use trial and error with many designed sequences until the desired level of response is obtained (Kosuri et al., 2013). Previous work however has also shown that rather than rely on such trial and error approaches, it also is possible to use thermodynamic models of regulation to accurately predict specific input-output characteristics (Bintu et al., 2005; Garcia and Phillips, 2011). Such models also provide non-obvious insight into what characteristics can be designed. We have shown how we can use regulatory sequence, through the design of specific LacI binding site sequences, to further control the level of gene expression. We can use a similar method to tune the RNAP sequence or any future transcription factor which is analyzed with Sort-Seq.

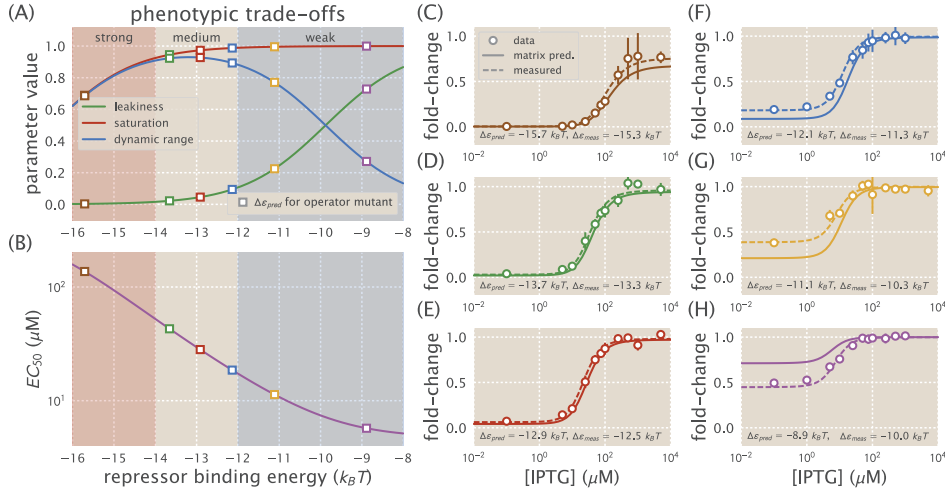


Figure 2.6: Energy matrix predictions can be used to design precise phenotypic responses. (A) Phenotypic parameters (leakiness, saturation, and dynamic range) exhibit trade-offs as $\Delta\epsilon_R$ is varied. Maximizing saturation or minimizing leakiness can only be achieved by reducing the dynamic range below its maximum. (B) Operators with different values of $\Delta\epsilon_R$ were chosen to have varying induction responses based on the phenotypic trade-offs shown in Part A. The induction responses predicted based on energy matrix predictions (solid lines) generally agree well with IPTG induction data obtained for each of the binding sites in a background strain with $R = 260$.

As a future step, we were interested in whether our sequence-energy mapping could be used to precisely design different induction responses. Induction is well described by the Monod-Wyman-Changeux (MWC) model (Monod, Wyman, and Changeux, 1965), with LacI in equilibrium between two conformations, termed the inactive and active states. In our formulation of fold-change as a function of inducer concentration, given in Eq 2.6, $p_A(c)$ is well described by

$$p_A(c) = \frac{(1 + \frac{c}{K_A})^2}{(1 + \frac{c}{K_A})^2 + e^{-\beta\Delta\epsilon_{AI}} (1 + \frac{c}{K_I})^2}, \quad (2.7)$$

where c is the concentration of inducer, K_A and K_I are the dissociation constants of the inducer and repressor when the repressor is in its active or inactive state, respectively, and $\Delta\epsilon_{AI}$ is the difference in free energy between the repressor's active and inactive states. Many of the parameters in Eq. 2.7 can and have been independently measured. Specifically, $K_A = 139 \mu M$, $K_I = 0.53 \mu M$, and $\Delta\epsilon_{AI} = -4.5 k_B T$.

We note that an induction response can be described by four key phenotypic pa-

rameters. The leakiness is the minimum fold-change when no inducer is present, given by fold-change($c \rightarrow 0$) from Eq. 2.6. The saturation is the maximum fold-change when inducer is present at saturating concentrations, given by fold change ($c \rightarrow \infty$). The dynamic range is the difference between the saturation and leakiness, and represents the magnitude of the induction response. Figure 2.6 (A) shows how these three phenotypic parameters vary with $\Delta\epsilon_R$ given the values of K_A , K_I , and $\Delta\epsilon_{AI}$ listed above and the repressor copy number $R = 260$. Lastly in Fig. 2.6 (B), the $[EC_{50}]$ of an induction response denotes the inducer concentration required to generate a response that is halfway between the minimum and maximum values.

There is an inherent trade-off between phenotypic parameters. For instance, tuning $\Delta\epsilon_P$ to be comparatively strong ($-8k_bT$), will increase the leakiness significantly. Mutating the DNA can adjust $\Delta\epsilon_P$ and $\Delta\epsilon_R$, while to adjust K_A or K_I the protein itself must be mutated.

To show how energy matrices can be used to design specific induction responses, we used the phenotypic trade-offs shown in Fig. 2.6 (A) to choose four different values of $\Delta\epsilon_R$ that would provide distinct outputs. These values were $\Delta\epsilon_R \approx -16 k_bT$, which would provide a minimal leakiness level but not reach full saturation; $\Delta\epsilon_R \approx -13 k_bT$, which would maximize dynamic range; $\Delta\epsilon \approx -11.5 k_bT$, which would maximize saturation but have an intermediate dynamic range; and $\Delta\epsilon_R \approx -10 k_bT$, which is close to the threshold between specific binding and nonspecific binding, and would provide a narrow dynamic range. Four of the single base-pair mutants designed in the previous section had predicted binding energies that matched these approximate values. Induction responses for each of the mutants were determined by growing cultures in the presence of varying IPTG concentrations and measuring the fold-change at each concentration. Fig 2.6 (B) shows how the induction data compare against fold-change curves plotted using $\Delta\epsilon_R$ values predicted from the energy matrix, and fold-change as defined in Eq. 2.1 and Eq. 2.7. The measured induction responses were found to match the theoretical predictions quite well, though for the sequence with a predicted energy of $\Delta\epsilon_R \approx -11.5k_bT$, we find that the $[EC_{50}]$ is shifted toward a higher IPTG concentration. This is at least in part due to a higher measured binding energy ($-12.5 k_bT$ instead of $-11.5 k_bT$) than predicted by our LacI energy matrix.

Sort-Seq can be used to probe both the DNA and amino acid interactions

So far we have examined how energy matrices provide us with a quantitative mapping between DNA sequence and binding energy, and how this can allow us to predict specific input-output characteristics. In this final section we show how we can also use energy matrices to investigate the effects of amino acid mutations on a transcription factor's sequence specificity. Specifically, we make individual amino acids changes to the repressor's DNA-binding domain and through additional Sort-Seq experiments, observe how those mutations modify the LacI energy matrix. This approach in particular makes it possible to determine how changing the amino acid composition of the DNA-binding domain alters DNA sequence preference.

We performed Sort-Seq using strains containing one of three LacI mutants, Y20I, Q21A, or Q21M, where the first letter indicates the wild-type amino acid, the number indicates the amino acid position, and the last letter indicates the identity of the mutated amino acid. These mutants have previously been found to alter LacI-DNA binding properties without entirely disrupting the repressor's ability to bind DNA (Milk, Daber, and Lewis, 2010; Daber and Sochor, 2011). We note that we use a slightly different version of LacI from the one used in Refs. (Milk, Daber, and Lewis, 2010; Daber and Sochor, 2011), so that the residue numbers in our version of LacI are shifted upward by 3 bp.

Sequence logos for each LacI mutant are shown in Figure 2.7, along with the wild-type sequence logo for comparison. As with the wild-type repressor, for each of the mutant repressors we find that the left half-site of the sequence logo has a higher information content. For both Y20I and Q21M, the same sequence is preferred in the left half-site as the wild-type sequence logo. This contrasts with the results from Milk, Daber, and Lewis, 2010, in which it was found that Y20I prefers an adenine at sequence position 7, rather than the guanine preferred at this position by the wild-type repressor. As in Milk, Daber, and Lewis, 2010, we find that an adenine is preferred at sequence position 8 for the Q21A mutant.

Some more subtle features can be observed when comparing the right half-sites. Within the right half-site, the most important base positions consistently appear to be 12, 13, 16, and 17. All mutants, along with the wild-type repressor, prefer cytosine and adenine at sequence positions 16 and 17. The wild-type, Q21A, and Q21M mutants all prefer an adenine and a tyrosine at positions 12 and 13, while the Y20I mutant prefers tyrosine and cytosine. For all mutants, the preferred bases at positions 16 and 17 are symmetrical to the corresponding bases in the left half-

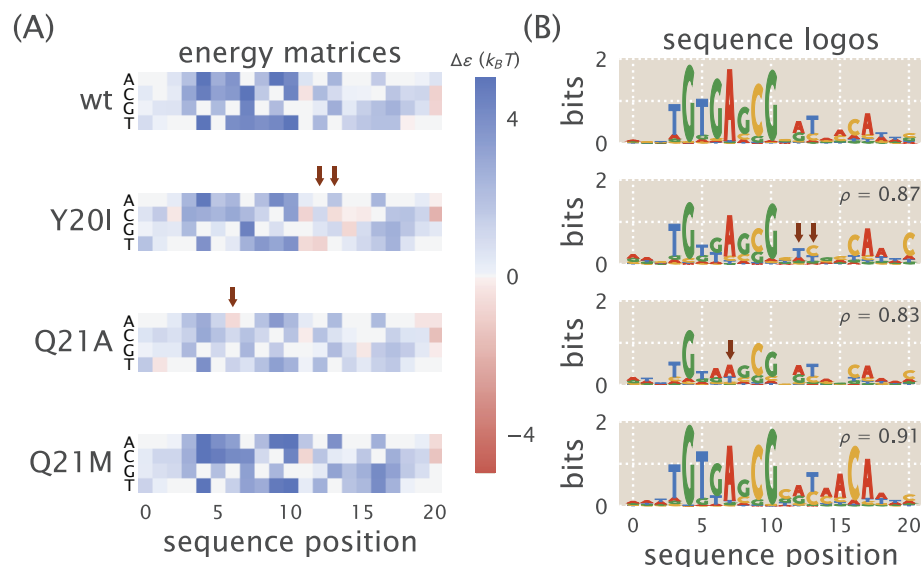


Figure 2.7: Point mutations to LacI DNA-binding domain cause subtle changes to sequence specificity. Mutations were made to residues 20 and 21 of LacI, both of which lie within the DNA-binding domain. The mutations Y20I and Q21A weaken the repressor-operator binding energy, while the mutation Q21M strengthens the binding energy. Y20I exhibits minor changes to specificity in low-information regions of the binding site, and Q21A experiences a change to specificity within a high-information region of the binding site. Specifically, Q21A prefers A at operator position 7 while the wild-type repressor prefers G at this position.

site (positions 4 and 5). By contrast, position 12 is consistently not symmetrical to position 8 in the right half-site, and position 13 for Y20I is not symmetrical to position 7 in the right half-site. Thus we see that the lac repressor's notable preference for a pseudo-symmetric binding site is preserved in each of the mutants we tested.

2.3 Discussion

We have shown how the massively parallel reporter assay, Sort-Seq (Kinney and Callan, 2010), can be used to generate a mapping between regulatory sequence and transcription factor binding energy using linear energy matrix models. By using a simple thermodynamic model, we find that this mapping provides further control over the input-output gene expression characteristics through finer control of the LacI DNA-binding energy. This work follows from a previous effort in our group to test the validity of such energy matrix models that describe binding of RNAP (Brewster, Jones, and Phillips, 2012). Here we explore whether the approach can be applied more broadly to other regulatory components. Specifically, we first used Sort-Seq

to map sequence to binding energy by inferring energy matrices for the repressor LacI. We perform this work in the context of a simple repression architecture, which represents a widespread bacterial regulatory architecture (Rydenfelt et al., 2014) that is commonly employed in synthetic biology (Brophy and Christopher A. Voigt, 2014; Khalil and Collins, 2010; Purnick and Weiss, 2009). We then demonstrate the validity of our model by designing roughly 30 mutant LacI binding site sequences, where we then demonstrate control over fold-change in gene expression, and show how such regulatory sequences can be used to optimize the inducible response of LacI by IPTG. Lastly, we show how Sort-Seq can also be used to probe the amino acid-DNA interactions. Here we perform Sort-Seq in several *E. coli* strains containing mutant LacI proteins and find only minor perturbations to the LacI sequence specificity following single amino-acid changes to the LacI DNA-binding domain.

While we focused on the regulatory component of LacI, we believe it will be possible to use regulatory sequence to predict gene expression more broadly across the bacterial genome and to other synthetic regulatory constructs, assuming that a thermodynamic model is in hand that can adequately describe the regulatory architecture. It is clear from our work that although we could accurately design regulatory sequences with a predictable fold-change, there were a variety of instances with notable discrepancies between the measured and predicted fold-change. This may suggest the need to consider more complex models than our linear energy matrices that incorporate non-additive contributions (Benos, Bulyk, and Gary D. Stormo, 2002). Deep-learning algorithms may provide an alternative approach to model the DNA-protein interactions (Sun et al., 2017). Future work on applying neural networks is discussed in B.9. Another consideration is that while Sort-Seq was performed on plasmids, our designed promoters were integrated on the chromosome, and aspects related to chromosomal context and DNA compaction are not considered in our model. Landing pad technologies for chromosomal integration (Kuhlman and Cox, 2010; Zhang et al., 2016; St-Pierre et al., 2013) could enable massively parallel reporter assays to be performed on chromosomes instead of on plasmids, and enable more accurate descriptions of chromosomally integrated promoters. Even when predicted fold-change did not match the observed fold-change, we still find a clear correlation between the predicted and measured LacI binding energies, and we have shown how regulatory sequence and a thermodynamic model can be used to guide our design of optimized inducible regulatory systems.

2.4 Methods

Sort-Seq libraries

To generate promoter libraries for Sort-Seq, mutagenized oligonucleotide pools were purchased from Integrated DNA Technologies (Coralville, IA). These consisted of single-stranded DNA containing the *lacUV5* promoter and LacI operator plus 15 bp on each end of PCR amplification. Either the *lacUV5* promoter and LacI binding site, or only the LacI binding site was mutated with a ten percent mutation rate per nucleotide. These oligonucleotides were amplified by PCR and inserted back into the pZS25-operator-YFP construct using Gibson Assembly. This plasmid is maintained in low copy (5-10 copies per cell) with the SC101 origin of replication (Lutz, 1997). To achieve high transformation efficiency, reaction buffer components from the Gibson Assembly reaction were removed by drop dialysis and cells were transformed by electroporation of freshly prepared cells. Following an initial outgrowth in SOC media, cells were diluted with 50 mL LB media and grown overnight under kanamycin selection. Transformation typically yielded $10^6 - 10^7$ colonies and were assessed by plating 100 μ L of cells diluted 1 : 10^4 onto an LB plate containing kanamycin.

DNA Constructs for fold-change measurements

Simple repression motifs used in Sort-Seq experiments and fold-change measurements were adapted from those in Garcia and Phillips, 2011. Briefly, the LacI operator (O1, O2, or O3) and YFP reporter gene were cloned into a pZS25 background directly downstream of a *lacUV5* promoter, driving expression of the YFP gene where the operator is not bound by LacI. This plasmid contains a kanamycin resistance gene for selection. Mutant LacI operator constructs were generated by PCR amplification of the *lacUV5* O1-YFP plasmid using primers containing the point mutations as well as sufficient overlap for re-circularizing the amplified DNA by Gibson Assembly.

A second construct was generated to provide expression of *lacI* gene. Here, *lacI* was cloned into a pZS3*1 background that provides constitutive expression of LacI from a $P_{LtetO-1}$ promoter (Lutz, 1997). This plasmid contains a chloramphenicol resistance gene for selection. To produce strains with different mean copy number of LacI that differ from the wild-type value of about 11 tetramers per cell, the ribosomal binding site for the *lacI* gene was mutated as described in (Salis, Mirsky, and Christopher A Voigt, 2009) using site-directed mutagenesis (Quickchange II; Stratagene, San Diego, CA) and further detailed in (Garcia and Phillips, 2011).

Bacterial Strains

E. coli strains used in this work were derived from K-12 MG1655. To generate strains with different LacI copy number, the *lacI* constructs were integrated into a strain that additionally has the entire *lacI* and *lacZYA* operons removed from the chromosome. These were integrated at the *ybcN* chromosomal location. This resulted in strains containing mean LacI copy numbers of $R = 30, 62, 130, 610, 870$, which were measured previously by quantitative western blots (Garcia and Phillips, 2011).

For Sort-Seq experiments, plasmid promoter libraries were constructed as described below and then transformed into the strains with different LacI copy number. For fold-change measurements, only the native O1 operator and associated mutants were considered. These simple repression constructs were chromosomally integrated at the *galK* chromosomal location. Generation of the final strains containing a simple repression motif and a specific LacI copy number was achieved by P1 transduction. For each LacI titration experiment, we also generated a strain where the entire *lacI* and *lacZYA* operons were removed, but with only the operator-YFP construct integrated. This provided us with a fluorescence expression measurement corresponding to $R = 0$, which is necessary for calculation of fold-change.

Sort-Seq fluorescence sorting

For each Sort-Seq experiment, cells were grown to saturation in lysogeny broth (LB) and then diluted 1 : 10,000 into minimal media (M9 + 0.5% glucose) for overnight growth. Once these cultures reached an OD 0.2-0.3 the cells were washed three times with PBS by centrifugation at 4000 rpm for 10 minutes and at 4°C. They were then diluted two-fold with PBS to reach an approximate OD of 0.1-0.15. These cells were then passed through a 40 μm cell strainer to eliminate any large clumps of cells.

A Beckman Coulter MoFlo XDP cell sorter was used to obtain initial fluorescence histograms of 500,000 events per library, which were used to set four binning gates that each covered 15% of the histogram. During sorting of each library, 500,000 cells were collected into each of the four bins. Finally, sorted cells were regrown overnight in 10 mL of LB media, under kanamycin selection.

Sort-Seq sequencing and data analysis

Overnight cultures from each sorted bin were miniprepmed (Qiagen, Germany), and PCR was used to amplify the mutated region from each plasmid for Illumina

sequencing. The primers contained Illumina adapter sequences as well as barcode sequences that enable pooling of the sorted samples. Sequencing was performed by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech or NGX Bio (San Francisco, CA). Single-end 100 bp or paired-end 150 bp flow cells were used, with about 500,000 sequences whose PHRED score was greater than 20 for each base pair, the total number of useful reads per bin was approximately 300,000 to 500,000 per million reads requested. Energy weight matrices for binding by LacI and RNAP were inferred using Bayesian parameter estimation with a error-model-averaged likelihood as previously described (Kinney and Callan, 2010; Kinney and Atwal, 2014).

Fold-change measurements by flow cytometry

Fold-change measurements were collected as previously described (Razo-Mejia et al., 2018) on a MACSquant Analyzer 10 Flow Cytometer (Miltenyi Biotec, Germany). Briefly, YFP fluorescence measurements were collected using 488nm laser excitation, with a 525/50 nm emission filter. Settings in the instrument panel for the laser were as follows: trigger on FSC (linear, 423V), SSC (linear, 537 V), and B1 laser (hlog, 790V). Before each experiment the MACSquant was calibrated using MACSquant Calibration Beads (Miltenyi Biotec, CAT NO. 130-093-607). Following growth of cells to OD 0.2-0.3, they were diluted ten fold in ice-cold minimal media (M9 + 0.5% glucose). Cells were then automatically sampled from a 96-well plate kept at approximately 4°C - 10°C using MACS Chill 96 Rack (Miltenyi Biotec, CAT NO. 130-094-459) at a flow rate of 2,000 - 6,000 measurements per second.

The fold-change in gene expression was calculated by taking the ratio of the mean YFP expression of the population of cells in the presence of LacI repressor to that in the absence of LacI repressor. Since the measured fluorescence intensity of each cell also includes autofluorescence which is present even in the absence of YFP, we account for this background by computing the fold change as

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{auto} \rangle}{\langle I_{R=0} \rangle - \langle I_{auto} \rangle} \quad (2.8)$$

where $\langle I_{R>0} \rangle$ is the average cell YFP intensity in the presence of repressor, $\langle I_{R=0} \rangle$ is the average cell YFP intensity in the absence of repressor, and $\langle I_{auto} \rangle$ is the average cell autofluorescence intensity.

Data curation

All data was collected, stored, and preserved using the Git version control software in combination with off-site storage and hosting website GitHub at url https://github.com/RPGroup-PBoC/seq_mapping. Sequencing data is available through the NCBI website under accession number SAMN08930313.

Acknowledgements

Access to the Miltenyi Biotec MACSquant Analyzer 10 Flow Cytometer was graciously provided by the Pamela Björkman lab at Caltech. We thank David Tirrell, Bradley Silverman, and Seth Lieblich for access and training for use of their Beckman Coulter MoFlo XDP cell sorter.

Supplemental Information: Summary of designed O1 binding site mutant

Mutated binding sites were created for the O1 binding site with 1, 2 or 3 mutations. All the designed sites are listed in Table 2.2. Each of these sequences have three predicted energies listed. Sort-seq matrices were generated starting from mutated libraries based on the WT sequences for the O1, O2, and O3 binding sites. There are 8 mutations between the O1 and O3 wild type sequences, and as such, these two generating libraries are very distant in sequence space. We can see that the predicted energies are different, but are generally within a few k_bT .

Identifier	LacI binding site sequence	O1 matrix prediction	O2 matrix prediction	O3 matrix prediction
mut005	AATTGTGAGCGGAGAACAATT	-11.929881	-13.428262	-12.243772
mut007	AATTGTGAGCGCATAACAATT	-15.633221	-14.197103	-15.296422
mut008	AATTGTGAGCGGATCACAATT	-15.520049	-14.133914	-14.986353
mut009	AATTGTGAGCGGAAAACAATT	-11.459789	-12.924778	-12.498838
mut010	AATTGCGAGCGGATAACAATT	-9.968247	-11.878477	-11.299124
mut011	AATTGTGAGGGGATAACAATT	-12.230209	-13.455658	-12.344994
mut012	AATTGTGAGCGGATATCAATT	-12.787483	-13.642761	-12.996080
mut013	AATTGTGAGCAGATAACAATT	-9.760610	-12.692912	-10.091807
mut014	AATTGTGAGAGGATAACAATT	-6.331624	-8.997448	-10.615486
mut102	AATTGTGAGCGGGTAACAATT	-13.641728	-13.896787	-14.788271
mut103	AAATGTGAGCGGATAACAATT	-13.328345	-13.584199	-14.401196
mut104	AATTGTGAGCGAGTAACAATT	-14.044856	-14.070952	-15.122752
mut105	ATTTGTGAGCGGAGAACAATT	-11.911801	-13.428375	-11.523189
mut107	CATTGTGAGCGCATAACAATT	-15.302753	-14.016493	-14.797621
mut108	AATTGTGAGCGGAACACAATT	-11.679837	-12.712688	-13.305983
mut109	AATTGTGAGCGGAATACAATT	-9.647010	-12.138189	-12.030819
mut111	AATTGTGAGGGGATAACAATC	-14.118290	-14.046511	-12.149832
mut201	AAATGTGAGCGAGTAACAATT	-13.558126	-13.874477	-14.571139
mut204	AATTGTGAGCGAGTAACAATT	-12.559931	-13.505622	-14.673368
mut205	ATTTGTGAGCGAAGAACAATT	-10.830003	-13.037210	-10.827536
mut207	CATTGTGAGCGCATAACATTT	-15.171401	-14.057285	-14.182531
mut208	AATTGTGAGCGGAACACAATG	-12.337016	-13.053090	-12.175545
mut209	AATTGTGAGCGGGATAACAATT	-9.473663	-12.254301	-11.857128
mut210	AATTGCGAGCGGATAACAAAG	-11.139112	-11.827513	-10.621422
mut211	AATTGTGAGGGTATAACAATC	-13.464516	-13.934262	-11.784251

Table 2.2: Summary of all energy predictions for mutant constructs.

We make these predictions as the average of LacI energy matrix created from Sort-Seq experiments where the mutated libraries are generated from either the O1, O2, or O3 wild type binding site sequences.

BIBLIOGRAPHY

- Ackers, Gary K and Alexander D Johnson (1982). “Quantitative model for gene regulation by A phage repressor”. en. In: *Proc. Natl. Acad. Sci. USA*, p. 5.
- Alekshun, M N and Levy (Oct. 1997). “Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon.” In: *Antimicrobial Agents and Chemotherapy* 41.10, pp. 2067–2075.
- Benos, Panayiotis V., Martha L. Bulyk, and Gary D. Stormo (Oct. 2002). “Additivity in protein–DNA interactions: how good an approximation is it?” en. In: *Nucleic Acids Research* 30.20, pp. 4442–4451. DOI: 10.1093/nar/gkf578.
- Berg, O. G. and P. H. von Hippel (1987). “Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters”. In: *J Mol Biol* 193.4, pp. 723–50. DOI: 10.1016/0022-2836(87)90354-8.
- Berger, Michael F. et al. (Nov. 2006). “Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities”. en. In: *Nature Biotechnology* 24.11, pp. 1429–1435. DOI: 10.1038/nbt1246.
- Bintu, Lacramioara et al. (Apr. 2005). “Transcriptional regulation by the numbers: models”. en. In: *Current Opinion in Genetics & Development*. Chromosomes and expression mechanisms 15.2, pp. 116–124. DOI: 10.1016/j.gde.2005.02.007.
- Brewster, Robert C., Daniel L. Jones, and Phillips (Dec. 2012). “Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*”. en. In: *PLoS Computational Biology* 8.12. Ed. by Erik van Nimwegen, e1002811. DOI: 10.1371/journal.pcbi.1002811.
- Brophy, Jennifer A. N. and Christopher A. Voigt (May 2014). “Principles of genetic circuit design”. en. In: *Nature Methods* 11.5, pp. 508–520. DOI: 10.1038/nmeth.2926.
- Buchler, Nicolas E., Ulrich Gerland, and Terence Hwa (Apr. 2003). “On schemes of combinatorial transcription logic”. en. In: *Proceedings of the National Academy of Sciences* 100.9, pp. 5136–5141. DOI: 10.1073/pnas.0930314100.
- Christensen, Ryan G. et al. (July 2011). “A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity”. en. In: *Nucleic Acids Research* 39.12, e83–e83. DOI: 10.1093/nar/gkr239.
- Daber, Robert and Mitchell Sochor Matthew and Lewis (May 2011). “Thermodynamic Analysis of Mutant lac Repressors”. en. In: *Journal of Molecular Biology*. The Operon Model and its Impact on Modern Molecular Biology 409.1, pp. 76–87. DOI: 10.1016/j.jmb.2011.03.057.
- Djordjevic, M. (Nov. 2003). “A Biophysical Approach to Transcription Factor Binding Site Discovery”. en. In: *Genome Research* 13.11, pp. 2381–2390. DOI: 10.1101/gr.1271603.

- Dror, Iris et al. (Sept. 2015). “A widespread role of the motif environment in transcription factor binding across diverse protein families”. en. In: *Genome Research* 25.9, pp. 1268–1280. doi: 10.1101/gr.184671.114.
- Fields, Dana S et al. (Aug. 1997). “Quantitative specificity of the Mnt repressor 11Edited by K. Yamamoto”. en. In: *Journal of Molecular Biology* 271.2, pp. 178–194. doi: 10.1006/jmbi.1997.1171.
- Foreman-Mackey et al. (Mar. 2013). “emcee: The MCMC Hammer”. In: *Publications of the Astronomical Society of the Pacific* 125.925. arXiv: 1202.3665, pp. 306–312. doi: 10.1086/670067.
- Gama-Castro, Socorro et al. (Jan. 2016). “RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond”. en. In: *Nucleic Acids Research* 44.D1, pp. D133–D143. doi: 10.1093/nar/gkv1156.
- Garcia and Phillips (July 2011). “Quantitative dissection of the simple repression input-output function”. en. In: *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–12178. doi: 10.1073/pnas.1015616108.
- Garcia, Alvaro Sanchez, et al. (July 2012). “Operator Sequence Alters Gene Expression Independently of Transcription Factor Occupancy in Bacteria”. en. In: *Cell Reports* 2.1, pp. 150–161. doi: 10.1016/j.celrep.2012.06.004.
- Grainger, David C. et al. (Dec. 2005). “Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome”. en. In: *Proceedings of the National Academy of Sciences* 102.49, pp. 17693–17698. doi: 10.1073/pnas.0506687102.
- Ireland, William T. and Kinney (May 2016). “MPAthic: Quantitative Modeling of Sequence-Function Relationships for massively parallel assays”. en. In: doi: 10.1101/054676.
- Jolma, Arttu et al. (Jan. 2013). “DNA-Binding Specificities of Human Transcription Factors”. en. In: *Cell* 152.1, pp. 327–339. doi: 10.1016/j.cell.2012.12.009.
- Jung, Cheulhee et al. (June 2017). “Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips”. en. In: *Cell* 170.1, 35–47.e13. doi: 10.1016/j.cell.2017.05.044.
- Khalil, Ahmad and James J. Collins (May 2010). “Synthetic biology: applications come of age”. en. In: *Nature Reviews Genetics* 11.5, pp. 367–379. doi: 10.1038/nrg2775.
- Kinney and Atwal (Jan. 2014). “Parametric Inference in the Large Data Limit Using Maximally Informative Models”. In: *Neural Computation* 26.4, pp. 637–653. doi: 10.1162/NECO_a_00568.
- Kinney and Callan (May 2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. en. In: *Proceedings of the National Academy of Sciences* 107.20, pp. 9158–9163. doi: 10.1073/pnas.1004290107.

- Kosuri, Sriram et al. (Aug. 2013). “Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*”. en. In: *Proceedings of the National Academy of Sciences* 110.34, pp. 14024–14029. doi: 10.1073/pnas.1301301110.
- Kuhlman and Edward C. Cox (Apr. 2010). “Site-specific chromosomal integration of large synthetic constructs”. en. In: *Nucleic Acids Research* 38.6, e92–e92. doi: 10.1093/nar/gkp1193.
- Levo, Michal, Tali Avnit-Sagi, et al. (Feb. 2017). “Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays”. en. In: *Molecular Cell* 65.4, 604–617.e6. doi: 10.1016/j.molcel.2017.01.007.
- Levo, Michal, Einat Zalckvar, et al. (July 2015). “Unraveling determinants of transcription factor binding outside the core binding site”. en. In: *Genome Research* 25.7, pp. 1018–1029. doi: 10.1101/gr.185033.114.
- Levy et al. (Oct. 2017). “A Synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial 54 Promoters”. en. In: *Cell Reports* 21.3, pp. 845–858. doi: 10.1016/j.celrep.2017.09.063.
- Lutz, R (Mar. 1997). “Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements”. en. In: *Nucleic Acids Research* 25.6, pp. 1203–1210. doi: 10.1093/nar/25.6.1203.
- Maerkl, and Quake (Jan. 2007). “A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors”. en. In: *Science* 315.5809, pp. 233–237. doi: 10.1126/science.1131007.
- Melnikov, Alexandre et al. (Mar. 2012). “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay”. en. In: *Nature Biotechnology* 30.3, pp. 271–277. doi: 10.1038/nbt.2137.
- Mendoza-Vargas, Alfredo et al. (Oct. 2009). “Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*”. en. In: *PLoS ONE* 4.10. Ed. by Chad Creighton, e7526. doi: 10.1371/journal.pone.0007526.
- Milk, Leslie, Robert Daber, and Mitchell Lewis (2010). “Functional rules for lac repressor–operator associations and implications for protein–DNA interactions”. en. In: *Protein Science* 19.6, pp. 1162–1172. doi: 10.1002/pro.389.
- Minchin, Stephen D. and Busby (Jan. 2009). “Analysis of mechanisms of activation and repression at bacterial promoters”. en. In: *Methods. Methods Related to Bacterial Transcriptional Control* 47.1, pp. 6–12. doi: 10.1016/j.ymeth.2008.10.012.

- Monod, Jacques, Jeffries Wyman, and Jean-Pierre Changeux (May 1965). “On the nature of allosteric transitions: A plausible model”. en. In: *Journal of Molecular Biology* 12.1, pp. 88–118. doi: 10.1016/S0022-2836(65)80285-6.
- Nutiu, Razvan et al. (June 2011). “Direct visualization of DNA affinity landscapes using a high-throughput sequencing instrument”. In: *Nature biotechnology* 29.7, pp. 659–664. doi: 10.1038/nbt.1882.
- Oehler et al. (Apr. 1990). “The three operators of the lac operon cooperate in repression.” en. In: *The EMBO Journal* 9.4, pp. 973–979. doi: 10.1002/j.1460-2075.1990.tb08199.x.
- St-Pierre, François et al. (Sept. 2013). “One-Step Cloning and Chromosomal Integration of DNA”. In: *ACS Synthetic Biology* 2.9, pp. 537–541. doi: 10.1021/sb400021j.
- Purnick, Priscilla E. M. and Ron Weiss (June 2009). “The second wave of synthetic biology: from modules to systems”. en. In: *Nature Reviews Molecular Cell Biology* 10.6, pp. 410–422. doi: 10.1038/nrm2698.
- Razo-Mejia, Manuel et al. (Apr. 2018). “Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction”. en. In: *Cell Systems* 6.4, 456–469.e10. doi: 10.1016/j.cels.2018.02.004.
- Rydenfelt, Mattias et al. (Dec. 2014). “The Influence of Promoter Architectures and Regulatory Motifs on Gene Expression in Escherichia coli”. en. In: *PLoS ONE* 9.12. Ed. by Jordi Garcia-Ojalvo, e114347. doi: 10.1371/journal.pone.0114347.
- Salis, Howard M, Ethan A Mirsky, and Christopher A Voigt (Oct. 2009). “Automated design of synthetic ribosome binding sites to control protein expression”. en. In: *Nature Biotechnology* 27.10, pp. 946–950. doi: 10.1038/nbt.1568.
- Shultzaberger, Ryan K. et al. (Mar. 2012). “Probing the Informational and Regulatory Plasticity of a Transcription Factor DNA–Binding Domain”. In: *PLoS Genetics* 8.3. doi: 10.1371/journal.pgen.1002614.
- Stormo, G. D. (Jan. 2000). “DNA binding sites: representation and discovery”. en. In: *Bioinformatics* 16.1, pp. 16–23. doi: 10.1093/bioinformatics/16.1.16.
- Sun, Tanlin et al. (May 2017). “Sequence-based prediction of protein protein interaction using a deep-learning algorithm”. In: *BMC Bioinformatics* 18.1, p. 277. doi: 10.1186/s12859-017-1700-2.
- Weinert, Franz M. et al. (Dec. 2014). “Scaling of Gene Expression with Transcription-Factor Fugacity”. In: *Physical Review Letters* 113.25, p. 258101. doi: 10.1103/PhysRevLett.113.258101.
- Weirauch, Matthew T et al. (Feb. 2013). “Evaluation of methods for modeling transcription factor sequence specificity”. en. In: *Nature Biotechnology* 31.2, pp. 126–134. doi: 10.1038/nbt.2486.

- Wunderlich, Zeba and Leonid A. Mirny (Oct. 2009). “Different gene regulation strategies revealed by analysis of binding motifs”. en. In: *Trends in Genetics* 25.10, pp. 434–440. doi: 10.1016/j.tig.2009.08.003.
- Xu, Denise J. and Marcus B. Noyes (Jan. 2015). “Understanding DNA-binding specificity by bacteria hybrid selection”. en. In: *Briefings in Functional Genomics* 14.1, pp. 3–16. doi: 10.1093/bfpg/elu048.
- Zhang, Huibin et al. (Apr. 2016). “Comprehensive mutagenesis of the fimS promoter regulatory switch reveals novel regulation of type 1 pili in uropathogenic *Escherichia coli*”. en. In: *Proceedings of the National Academy of Sciences* 113.15, pp. 4182–4187. doi: 10.1073/pnas.1522958113.

Chapter 3

A SYSTEMATIC APPROACH FOR DISSECTING THE MOLECULAR MECHANISMS OF TRANSCRIPTIONAL REGULATION IN BACTERIA.

A version of this chapter originally appeared as N. M. Belliveau, S. L. Barnes, W. T. Ireland, D. L. Jones, M. J. Sweredoski, A. Moradian, S. Hess, J. B. Kinney, R. Phillips (2018). A systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences*, <http://doi.org/10.1073/pnas.1722055115>.

Author contribution note: for this chapter, I (WI) assisted with experimental design, data analysis, and manuscript writing.

3.1 Introduction

The sequencing revolution has left in its wake an enormous challenge: the rapidly expanding catalog of sequenced genomes is far outpacing a sequence-level understanding of how the genes in these genomes are regulated. This ignorance extends from viruses to bacteria to archaea to eukaryotes. Even in *E. coli*, the model organism in which transcriptional regulation is best understood, we still have no indication if or how more than half of the genes are regulated (Fig 1.7; Gama-Castro et al., 2016; Keseler et al., 2013). In other model bacteria such as *Bacillus subtilis*, *Caulobacter crescentus*, *Bibrio harveyi*, or *Pseudomonas aeruginosa*, far fewer genes have established regulatory mechanisms (Munch, 2003; Cipriano et al., 2013; Kılıç et al., 2014).

New tools are needed for studying regulatory architecture in these and other bacteria. Although an arsenal of genetic and biochemical methods have been developed for dissecting promoter function at individual bacterial promoters (reviewed in Minchin and Busby, 2009), these methods are not readily parallelized. As a result, they will likely not lead to a comprehensive understanding of full regulatory genomes anytime soon. RNA sequencing, chromatin immunoprecipitation, and other high throughput techniques are increasingly being used to study gene regulation in *E. coli* (Grainger et al., 2005; Bonocora and J. T. Wade, 2015; Latif et al., 2018; Zheng et al., 2004; Singh et al., 2014; Vvedenskaya, Goldman, and Nickels, 2015;

THE REGULATORY GENOME OF *ESCHERICHIA COLI*: PROMOTER STUDIED

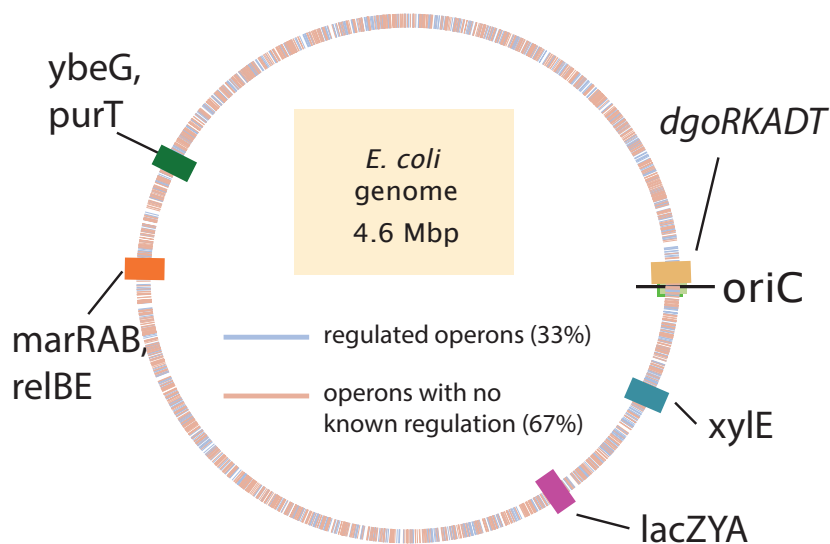


Figure 3.1: Summary of transcriptional regulatory knowledge in *E. coli*. left panel: Well-characterized promoters considered in this work. The schematics highlight the known regulatory architectures for the annotated promoters of *marRAB*, *relBE*, and *lacZYA*. The center plot identifies the genomic location of different operons in *E. coli*. Operons with annotated TF binding sites are shown in light blue, while those lacking regulatory descriptions are shown in light red (Gama-Castro et al., 2016). The genomic location of the promoters considered in this work are labeled.

Wade, 2005)), but these methods are incapable of revealing either the nucleotide resolution location of all functional transcription factor binding sites, or the way in which interactions between DNA-bound transcription factors and RNA polymerase modulate transcription.

In recent years a variety of massively parallel reporter assays have been developed for dissecting the functional architecture of transcriptional regulatory sequences in bacteria, yeast, and metazoans. These technologies have been used to infer biophysical models of well-studied loci, characterize synthetic promoters constructed from known binding sites, and search for new transcriptional regulatory sequences (Kinney, Murugan, et al., 2010; Melnikov et al., 2012; Kheradpour et al., 2013; Patwardhan et al., 2012; Sharon et al., 2012; Kosuri et al., 2013; Arnold et al.,

2013; Maricque, Dougherty, and Cohen, 2016). CRISPR assays have also shown promise for identifying longer range enhancer-promoter interactions in mammalian cells (Fulco et al., 2016). However, no approach for using massively parallel reporter technologies to decipher the functional mechanisms of previously uncharacterized regulatory sequences has yet been established.

Here we describe a systematic and scalable approach for dissecting the functional architecture of previously uncharacterized bacterial promoters at nucleotide resolution using a combination of genetic, functional, and biochemical measurements. First, a massively parallel reporter assay, Sort-Seq (Kinney and Callan, 2010) is performed on a promoter in multiple growth conditions in order to identify functional transcription factor binding sites. DNA affinity chromatography and mass spectrometry (Mittler, Butter, and M. Mann, 2008; Mirzaei et al., 2013) are then used to identify the regulatory proteins that recognize these sites. In this way one is able to identify both the functional transcription factor binding sites and cognate transcription factors in previously unstudied promoters. Subsequent massively parallel assays are then performed in gene-deletion strains to provide additional validation of the identified regulators. In many cases, the reporter data thus generated can further be used to infer quantitative models of transcriptional regulation.

In what follows, we first describe the application of this approach to four previously annotated promoters: *lacZYA*, *relBE*, *marRAB*, and *yebG*. This illustrates the overarching logic of our method and provides a benchmark for how well these methods work. We then describe this strategy applied to the previously uncharacterized promoters of *purT*, *xylE*, and *dgoRKADT*. These results demonstrate the ability to go from complete regulatory ignorance to an explicit quantitative model of a promoter's input-output behavior.

3.2 Results

To dissect how a promoter is regulated, we begin by performing Sort-Seq (Kinney and Callan, 2010). As shown in Fig 3.2, Sort-Seq works by first generating a library of cells, each of which contains a mutated promoter that drives expression of GFP from a low copy plasmid (5-10 copies per cell; Lutz, 1997) and provides a read-out of transcriptional state. We use fluorescence-activated cell sorting (FACS) to sort cells into multiple bins gated by their fluorescence level and then sequence the mutated plasmids from each bin. We found it sufficient to sort the libraries into four bins and generated data sets of approximately 0.5-2 million sequences across the sorted bins

(Fig. 3.12 (A)-(D)). Putative binding sites were identified by examining expression shift plots which show the average change in fluorescence when each position is mutated (Fig. 3.2(B)). Mutations to the DNA will disrupt binding of transcription factors, so regions with a positive shift are suggestive of binding by a repressor, while a negative shift suggests binding by an activator or RNA polymerase (RNAP).

The identified binding sites are further interrogated by performing information-based modeling with the Sort-Seq data. Here we generate energy matrix models (Kinney and Callan, 2010; Ireland and Kinney, 2016) that describe the sequence-dependent energy of interaction of a transcription factor at the putative binding site. For each matrix, we use a convention that the wild-type sequence is set to have an energy of zero (see example energy matrix in Fig. B.4. Mutations that enhance binding are identified in blue, while mutations that weaken binding are identified in red. We also use the energy matrices to generate sequence logos (Berg and Hoppel, 1987; Schneider and Stephens, 1990; Stormo, 2000) which provides a useful visualization of the sequence-specificity (see above matrix in Fig. 3.2(B)).

We next perform DNA affinity chromatography experiments using DNA oligonucleotides containing the binding sites identified by Sort-Seq. Here we apply a stable isotopic labeling of cell culture (SILAC) approach (Ong et al., 2002), which enables us to perform a second reference chromatography experiment that is simultaneously analyzed by mass spectrometry to identify the target transcription factor. As shown in Fig.3.2(C), we begin by preparing two cell lysates: one with cells supplemented with natural lysine and the other with a heavy isotopic form of lysine. We then perform chromatography using magnetic beads with the tethered oligonucleotides. Our reference experiment is performed identically, except that the binding site has been mutated away from the original sequence (and is performed using the light lysate). The abundance of each protein is determined by mass spectrometry and used to calculate protein enrichment ratios, with the target transcription factor expected to exhibit a ratio greater than one. Most proteins detected will exhibit a protein enrichment near one due to non-specific binding in both purifications.

The energy matrix models and results from each DNA affinity chromatography experiment provide insight into the identity of each regulatory factor and hypotheses about potential regulatory mechanisms. In some instances we are able to test these hypotheses further with additional information-based modeling of thermodynamic models on our Sort-Seq data. Finally, to confirm binding by an identified regulator we perform Sort-Seq experiments in gene deletion strains, which no longer show

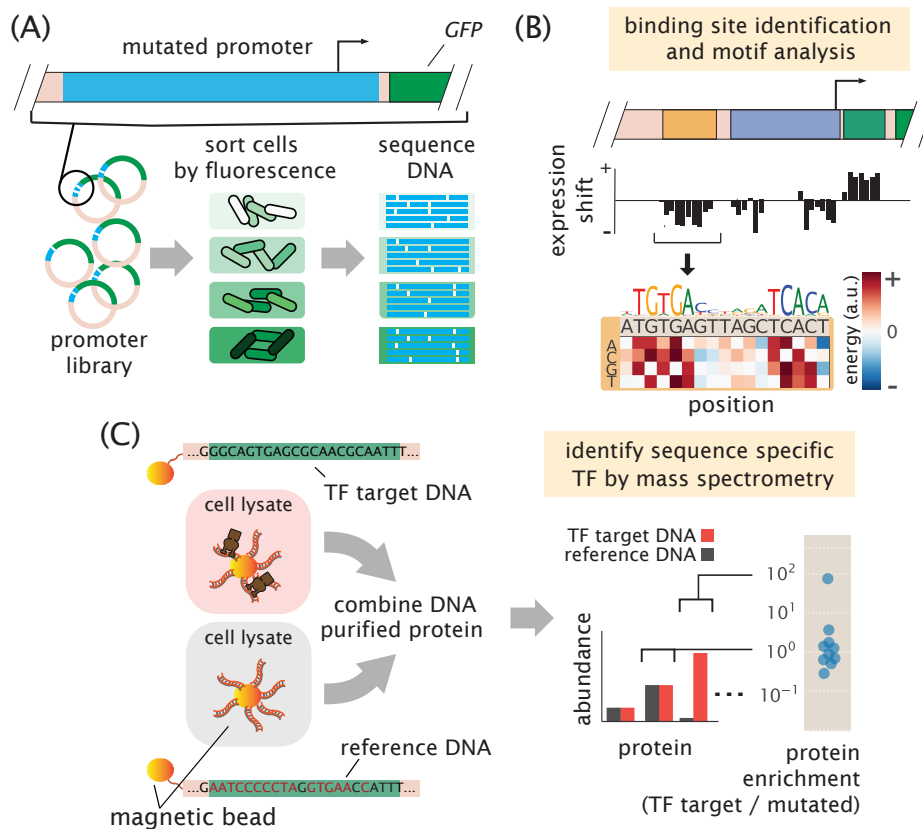


Figure 3.2: Overview of approach to characterize transcriptional regulatory DNA, using Sort-Seq and mass spectrometry. (A) Schematic of Sort-Seq. A promoter plasmid library is placed upstream of GFP and is transformed into cells. The cells are sorted into four bins by FACS and after regrowth, plasmids are purified and sequenced. The entire intergenic region associated with a promoter is included on the plasmid and a separate downstream ribosomal binding site sequence is used for translation of the GFP gene. (B) Regulatory binding sites are identified by calculating the average expression shift due to mutation at each position. The schematic shows the expression shift on a promoter region containing an activator (orange), RNAP (blue), and repressor (green) binding site. Quantitative models can be inferred to describe the associated DNA-protein interactions. An example energy matrix that describes the binding energy between an as yet unknown activator to the DNA is shown. By convention, the wild-type nucleotides have zero energy, with blue squares identifying mutations that enhance binding (negative energy), and where red squares reduce binding (positive energy). (C) Identify sequence specific TF by mass spectrometry. The schematic shows the process of identifying sequence specific TF by mass spectrometry. It involves cell lysate, TF target DNA, reference DNA, and a magnetic bead. The process includes combining DNA and purified protein, followed by mass spectrometry analysis. The resulting abundance of protein is shown on a log scale (10⁻¹ to 10²), and the protein enrichment (TF target / mutated) is shown on a log scale (10⁻¹ to 10²).

the positive or negative shift in expression along the binding site.

Sort-Seq recovers the known regulatory features of well-characterized promoters

To first demonstrate Sort-Seq as a tool to discover regulatory binding sites *de novo* we began by looking at the promoters of *lacZYA*, *relBE*, and *marRAB* (Oehler et al., 1990; Grainger et al., 2005; Alekshun and Levy, 1997). These promoters have been studied extensively and provide a useful test bed of distinct regulatory motifs to test our approach. To proceed we constructed libraries for each promoter by mutating their known regulatory binding sites. We also considered two different mutation frequencies in our libraries. For *lac*, our library had a mutation rate of approximately three percent per bp, while *mar* and *rel* had a rate of roughly nine percent per bp. For a 20 bp binding site, this corresponds to an average of less than one mutation per sequence at the low mutation rate, and about two mutations at the high mutation rate (See Supplemental Section 3.8 and Fig. 3.12(E),(F) for additional characterization).

We begin by considering the *lac* promoter. It contains three *lac* repressor (LacI) binding sites, two of which we consider here, and a cyclic AMP receptor (CRP) binding site. It exhibits the classic catabolic switch-like behavior that results in diauxie when *E. coli* is grown in the presence of glucose and lactose sugars (Loomis and Magasanik, 1967; Oehler et al., 1990; Busby and Ebright, 1999). We performed Sort-Seq with cells grown in M9 minimal media at 37°C. The information footprints and expression shifts at each position are shown in Fig. 3.3(A), with annotated binding sites from RegulonDB noted above the plot. The expression shifts reflect the expected regulatory role of each binding site, showing positive shifts for LacI and negative shifts for CRP and RNAP. The difference in magnitude at the two LacI binding sites likely reflect the different binding energies between these two binding site sequences, with LacI O3 having an *in vivo* dissociation constant that is almost three orders of magnitude weaker than the LacI O1 binding site (Oehler et al., 1990; Garcia and Phillips, 2011).

Next we consider the *rel* promoter that transcribes the toxin-antitoxin pair RelBE and RelB. It is one of about 36 toxin-antitoxin systems found on the chromosome, with important roles in cellular physiology including cellular persistence (Grainger et al., 2005; Yamaguchi and Inouye, 2011; Maisonneuve and Gerdes, 2014). When the toxin, RelE, is in excess of its cognate binding partner, the antitoxin RelB, the toxin causes cellular paralysis through cleavage of mRNA (Griffin, Davis, and Strobel,

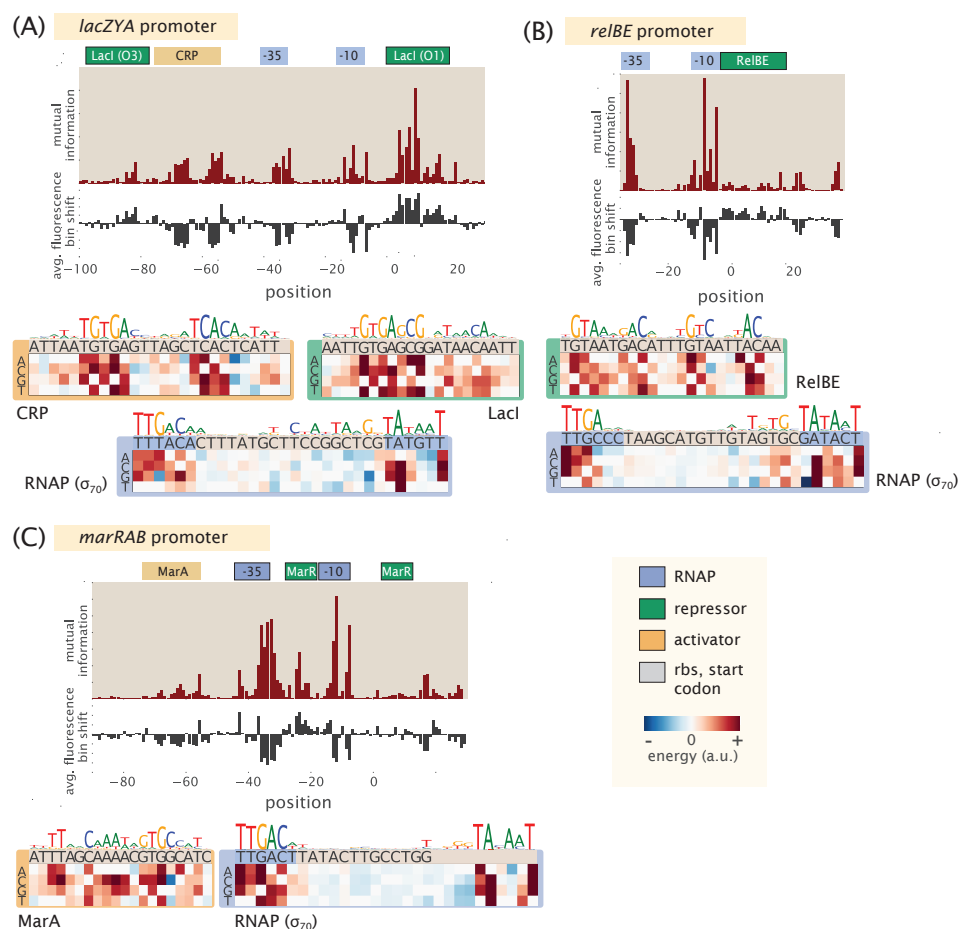


Figure 3.3: Sort-Seq identifies the regulatory landscape of the *lac*, *rel*, and *mar* promoters. (A) Sort-Seq of the *lac* promoter. Cells were grown in M9 minimal media with 0.5% glucose. Expression shifts are shown, with annotated binding sites for CRP (activator), RNAP (-10 and -35 subsites), and LacI (repressor) noted. Energy matrices and sequence logos are shown for each binding site. (B) Sort-Seq of the *rel* promoter. Cells were also grown in M9 minimal media with 0.5% glucose. The information footprints and expression shifts identify the binding sites of RNAP and RelBE (repressor), and energy matrices and sequence logos are shown for these. (C) Sort-Seq of the *mar* promoter. Here cells were grown in Lysogeny broth (LB) at 30°C. The expression shifts identify the known binding sites of Fis and MarA (activators), RNAP, and MarR (repressor). Energy matrices and sequence logos are shown for MarA and RNAP.

2013). Interestingly, the antitoxin protein also contains a DNA binding domain and is a repressor of its own promoter (Gotfredsen and Gerdes, 1998; Overgaard, Borch, and Gerdes, 2009; Cataudella, Trusina, et al., 2012; Cataudella, Sneppen, et al., 2013). We performed Sort-Seq with cells grown in M9 minimal media and at 37°C. The expression shifts are shown in Fig. 3.3(B) and were consistent with binding by RNAP and RelBE. In particular, a positive shift was observed at the binding site for RelBE, and the RNAP binding site showed mainly a negative shift in expression.

The third promoter, *mar*, is associated with multiple antibiotic resistance since its operon codes for the transcription factor MarA, which activates a variety of genes including the major multi-drug resistance efflux pump, ArcAB-tolC, and increases antibiotic tolerance (Aleksun and Levy, 1997). The *mar* promoter is itself activated by MarA, SoxS, and Rob (via the so-called marbox binding site), and further enhanced by Fis, which binds upstream of this marbox (Martin and Rosner, 1997). Under standard laboratory growth it is under repression by MarR (Aono, Tsukagoshi, and M. Yamamoto, 1998). We found that the promoter's fluorescence was quite dim in M9 minimal media and instead grew libraries in lysogeny broth (LB) at 30°C (Seoane and Levy, 1995). Again, the different features in the information footprint and expression shift plot (Fig. 3.3(C)) appeared to be consistent with the noted binding sites. One exception was that the downstream MarR binding site was not especially apparent. Both positive and negative expression shifts were observed along its binding site, which may be due to overlap with other features present including the native ribosomal binding site. There have also been reported binding sites for CRP (Ruiz and Levy, 2010; Zheng et al., 2004), Cra (Shimada, K. Yamamoto, and A. Ishihama, 2011), CpxR/CpxA (Weatherspoon-Griffin et al., 2014), and ArcR (Lee, Cho, and Kim, 2014). However these studies either required overexpression of the associated transcription factor, were computationally identified, or demonstrated through *in vitro* assays and were not observed under the growth condition considered here.

While each promoter qualitatively showed the expected regulatory behavior in each expression shift plot, we were also interested in whether we could recover the quantitative sequence specificity of each transcription factor from our data. We inferred energy matrices and associated sequence logos for the binding sites of RNAP, LacI, CRP, RelBE, MarA, and Fis. These are shown in Fig. 3.3 (A)-(C) and Fig. 3.4, and agreed with sequence logos generated from known genomic binding sites for these transcription factors (Pearson correlation coefficient $r = 0.5 - 0.9$;

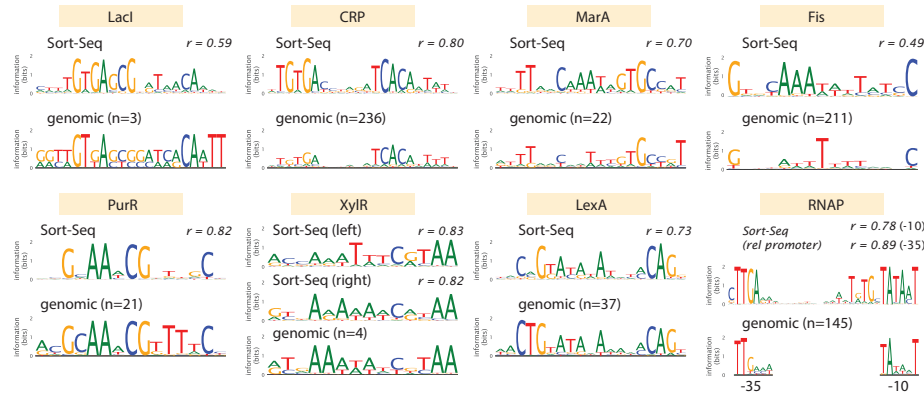


Figure 3.4: Comparison between Sort-Seq and genomic-based sequence logos. Comparisons are shown for LacI, CRP, MarA, Fis, PurR, XylR, LexA, and RNAP. Binding site sequences were obtained from RegulonDB, where n identifies the number of genomic binding sites that were used to construct the sequence logo. The Sort-Seq RNAP logo is based on data from the *rel* promoter. For the genomic RNAP logo, sequences were taken from computationally predicted RNAP binding sites on RegulonDB (top 3.3% scored sequences using their reported metric) for the 6 bp regions of the -10 and -35 binding sites. Pearson correlation coefficients are calculated with Equation 4.7 using the position weight matrices from the Sort-Seq and genomic matrices. For LexA, the first four bp were not used in the calculation due to overlap with the -10 RNAP binding site of the *yebG* promoter.

see Supplemental Section 3.9).

Identification of transcription factors with DNA affinity chromatography and quantitative mass spectrometry.

For our purpose of completely dissecting a promoter, it was next important to show that DNA affinity chromatography could indeed be used to identify transcription factors in *E. coli*. In particular, a challenge arises in identifying transcription factors due to their very low abundance. In *E. coli* the cumulative distribution in protein copy number shows that more than half have a copy number less than 100 per cell, with 90 % having copy number less than 1,000 per cell. This is several orders of magnitude below that of many other cellular proteins (Li et al., 2014).

We began by applying the approach to known binding sites for LacI and RelBE. For LacI, which is present in *E. coli* in about 10 copies per cell, we used the strongest binding site sequence, Oid (*in vivo* $K_d \approx 0.05nM$), and the weakest natural operator sequence, O3 (*in vivo* $K_d \approx 110nM$) (Oehler et al., 1990; S. Oehler, 2006; Kuhlman, Z. Zhang, et al., 2007; Garcia and Phillips, 2011). In Fig. 3.5(A) we plot the protein

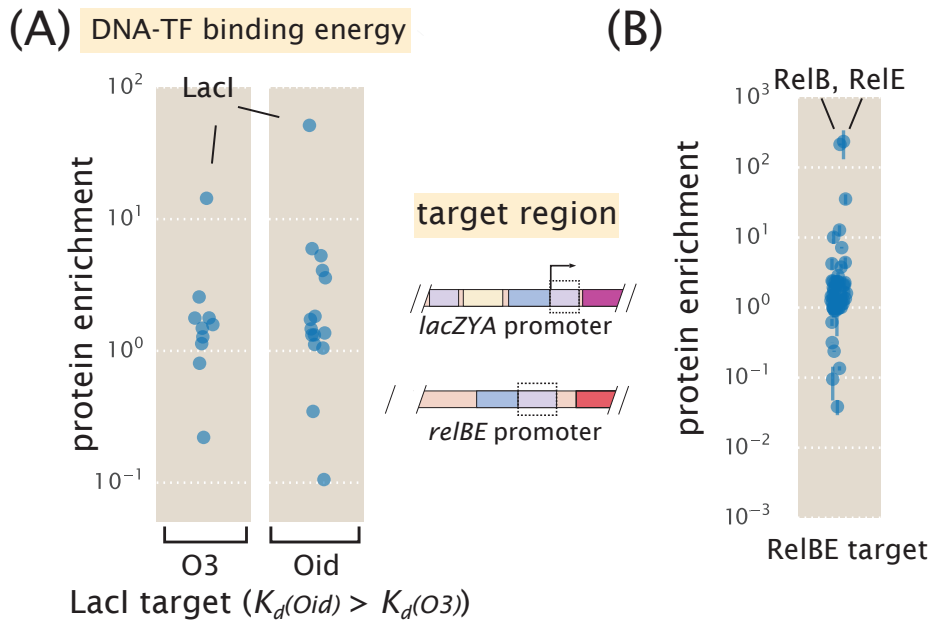


Figure 3.5: DNA affinity purification and identification of LacI and RelBE by mass spectrometry using known target binding sites. (A) Protein enrichment using the weak O3 binding site and strong synthetic Oid binding sites of LacI. LacI was the most significantly enriched protein in each purification. The target DNA region was based on the boxed area of the *lac* promoter schematic, but with the native O1 sequence replaced with either O3 or Oid. Data points represent average protein enrichment for each detected transcription factor, measured from a single purification experiment. (B) For purification using the RelBE binding site target, both RelB and its cognate binding partner RelE were significantly enriched. Data points show the average protein enrichment from two purification experiments. The target binding site is similarly shown by the boxed region of the *rel* promoter schematic. Data points in each purification show the protein enrichment for detected transcription factors. The gray shaded regions show where 95% of all detected protein ratios were found.

enrichments from each transcription factor identified by mass spectrometry. LacI was found with both DNA targets, with fold enrichment greater than 10 in each case, and significantly higher than most of the proteins detected (indicated by the shaded region, which represents the 95% probability density region of all proteins detected, including non-DNA binding proteins). Purification of LacI with about 10 copies per cell using the weak O₃ binding site sequence is near the limit of what would be necessary for most *E. coli* promoters.

To ensure this success was not specific to LacI, we also applied chromatography to the RelBE binding site. RelBE provides an interesting case since the strength of binding by RelB to DNA is dependent on whether RelE is bound in complex to RelB. There is at least a 100 fold weaker dissociation constant reported in the absence of RelE (G.-Y. Li et al., 2008; Overgaard, Borch, Jørgensen, et al., 2008). As shown in Fig. 3.5(B), we found over 100 fold enrichment of both proteins by mass spectrometry. As a consequence of performing a second reference purification, we find that fold enrichment should mostly reflect the difference in binding energy between the DNA sequences used in the two purifications, and be much less dependent on whether the protein was in low or high abundance within the cell. This appeared to be the case when considering other *E. coli* strains with LacI copy numbers between about 10 and 1,000 copies per cell (Fig. 3.6 (C)). Further characterization of the measurement sensitivity and dynamic range of this approach is noted in Supplemental Section 3.12.

Sort-Seq discovers regulatory architectures in unannotated regulatory regions.

Given that more than half of the promoters in *E. coli* have no annotated transcription factor binding sites in RegulonDB, we narrowed our focus by using several high throughput studies to identify candidate genes to apply our approach (Marbach et al., 2012; Schmidt et al., 2016). The work by Schmidt et al., 2016 in particular measured the protein copy number of about half the *E. coli* genes across 22 distinct growth conditions. Using this data, we identified genes that had substantial differential gene expression patterns across growth conditions, thus hinting at the presence of regulation and even how that regulation is elicited by environmental conditions (see further details in Supplemental Information Section A and Fig. 3.7(A)-(C)). On the basis of this survey, we chose to investigate the promoters of *purT*, *xylE*, and *dgoRKADT*. To apply Sort-Seq in a more exploratory manner, we considered three 60 bp mutagenized windows spanning the intergenic region of each gene. While it is certainly possible that regulatory features will lie outside of this window, a search

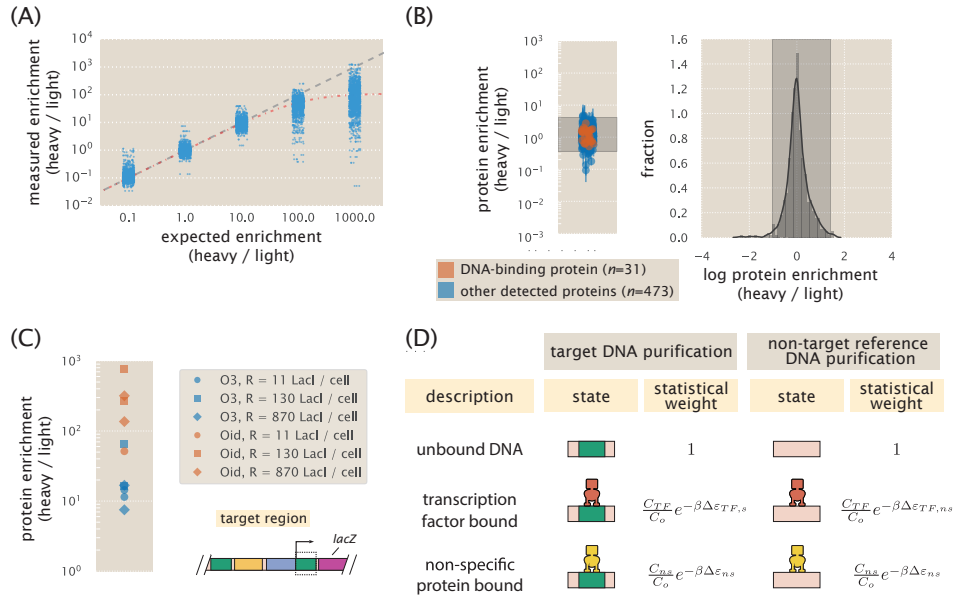


Figure 3.6: Identification of transcription factors using DNA-affinity chromatography and mass spectrometry. (A) Characterization of stable isotopic lysine labeling and mass spectrometry measurement sensitivity. Lysates from cell cultures grown in either heavy ($^{13}C_6^{15}N_2$ -L-lysine) or normal L-lysine were combined at ratios between 0.1:1 to 1000:1 heavy:light and the measured ratios in abundance are plotted for each protein. Note that for the 1:1 ratio we found a median ratio of 0.71. We therefore renormalized the ratio values using this as a correction factor. Data points represent the average values from $n = 3$ replicates. The gray line represents the expected measurement under perfect labeling, while the red line represents a 99.1% labeling efficiency (assuming that some fraction of heavy lysate is unlabeled). (B) DNA-affinity purification using the same DNA oligonucleotide to purify protein for both heavy and light cell lysates ($n = 3$). The scatter plot shows the average enrichment values for each protein detected. Proteins with DNA binding motifs (Keseler et al., 2013) are shown in red ($n = 41$), while other detected proteins are in blue ($n = 581$). Error bars represent the standard deviation, calculated from log protein enrichment values. The histogram shows the distribution of the measured ratios for all detected proteins, with 95% of the measurements contained between a log enrichment of -1.5 and 1.2, as indicated by the shaded region. (C) DNA-affinity purification of LacI using three different *E. coli* strains. Operator strength was varied by purifying LacI with either the weak O3 or strong Oid operators. LacI was detected as the most significantly enriched protein among all proteins detected. (D) States and weights are shown for an oligonucleotide in which a target transcription factor and other cellular proteins compete for a DNA binding site.

of known regulatory binding sites suggest that this should be sufficient to capture just over 70% of regulatory features in *E. coli* and provide a useful starting point (Fig. 3.7(D)).

The *purT* promoter contains a simple repression architecture and is repressed by PurR.

The first of our candidate promoters is associated with expression of *purT*, one of two genes found in *E. coli* that catalyze the third step in *de novo* purine biosynthesis (Rolfes, 2006; Cho et al., 2011). Due to a relatively short intergenic region, about 120 bp in length that is shared with a neighboring gene *yebG*, we also performed Sort-Seq on the *yebG* promoter (oriented in the opposite direction (Lomba et al., 1997); see schematic in Fig. 3.8(A)). To begin our exploration of the *purT* and *yebG* promoters, we performed Sort-Seq with cells grown in M9 minimal media with 0.5% glucose. The associated expression shift plots are shown in Fig. 3.8(A). While we performed Sort-Seq on a larger region than shown for each promoter, we only plot the regions where regulation was apparent.

For the *yebG* promoter, the features were largely consistent with prior work, containing a binding sites for LexA and RNAP. However, we found that the RNAP binding site is shifted 9 bp downstream from what was identified previously through a computational search (Lomba et al., 1997), demonstrating the ability of our approach to identify and correct errors in the published record. We were also able to confirm that the *yebG* promoter was induced in response to DNA damage by repeating Sort-Seq in the presence of mitomycin C (a potent DNA cross-linker known to elicit the SOS response and proteolysis of LexA (Wade, 2005); see Fig. 3.10(A), (B), and (D)).

Given the role of *purT* in the synthesis of purines, and the tight control over purine concentrations within the cell (Rolfes, 2006), we performed Sort-Seq of the *purT* promoter in the presence or absence of the purine or adenine, in the growth media. In growth without adenine (Fig. 3.8(A), right plot), we observed two negative regions in the information footprint and expression shift plots. We infer and energy matrix and examine the sequence preference of the site, these two features were identified as the -10 and -35 regions of an RNAP binding site. While these two features were still present upon addition of adenine, as shown in Fig. 3.8(B), this growth condition also revealed a putative repressor site between the -35 and -10 RNAP binding sites, indicated by a positive shift in expression (green annotation).

Following our strategy to find not only the regulatory sequences, but also their

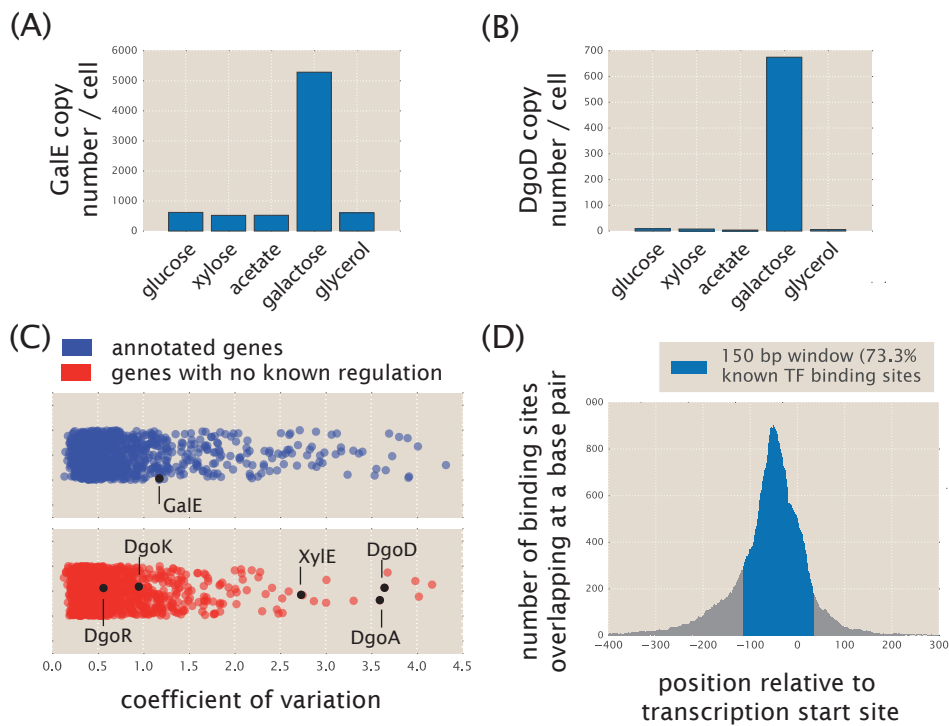


Figure 3.7: Identification of unannotated genes with potential regulation and distribution of known transcription factor binding sites in *E. coli*. (A) Here we show the protein copy numbers per cell for GalE across several carbon sources. Expression was sensitive to the presence of galactose which is consistent with its known regulation (with about 5000 copies per cell, versus about 500 for most other growth conditions). (B) DgoD was also found to be sensitive to the presence of galactose as the carbon source. The copy number was measured to be 675 copies per cell when cells were grown in galactose, and 15 copies per cell or less in all other conditions considered. For both (A) and (B), values are shown for growth in M9 minimal media, with glucose, xylose, acetate, galactose, and glycerol as carbon sources and obtained from (Schmidt et al., 2016). (C) Coefficient of variation (standard deviation divided by mean copy number) across the 22 growth conditions for each protein measured in (Schmidt et al., 2016). Proteins are identified as either having regulatory annotation (blue) or not (red) using the annotations in RegulonDB (Gama-Castro et al., 2016). GalE is noted among the annotated genes and provides a reference as a gene that is known to be regulated and be perturbed in this study, as shown in (A). (D) The histogram shows the genome-wide distribution of transcription factor binding sites relative to their respective transcription start sites. Binding sites were compiled from RegulonDB and used to calculate the number of overlapping binding sites at each position using the length and position of each binding site sequence. The location of the 150 bp mutation window used in this study is shown in blue, expected to capture upwards of 70% of known transcription factor binding site position.

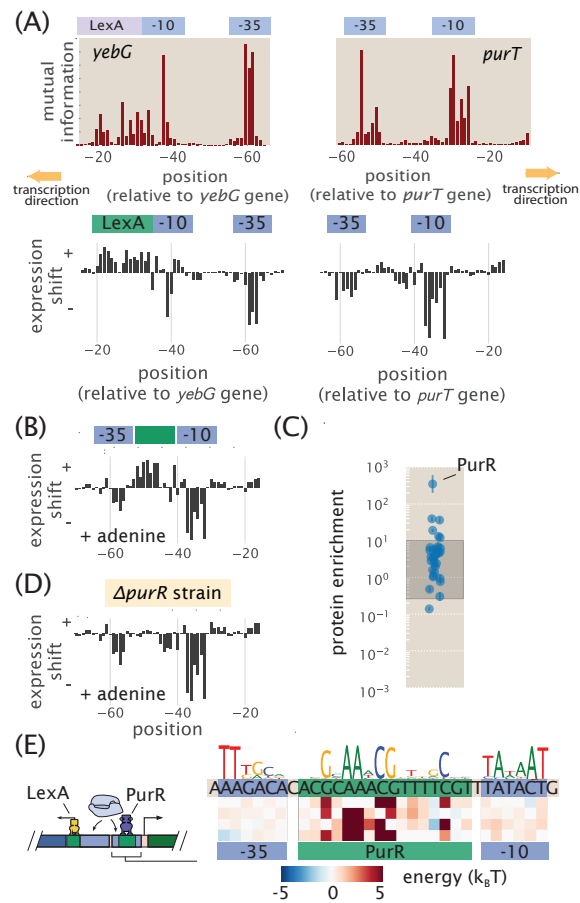


Figure 3.8: Sort-Seq distinguishes directional regulatory features and uncovers the regulatory architecture of the *purT* promoter. (A) A schematic is shown for the approximately 120 bp region between the *yebG* and *purT* genes, which code in opposite directions. Information footprints and expression shifts are shown for 60 bp regions where regulation was observed for each promoter, with positions noted relative to the start codon of each native coding gene. The -10 and -35 RNAP binding sites are identified in blue. (B) Expression shifts for the *purT* promoter, but in M9 minimal media with 0.5% glucose supplemented with adenine (100 μ g/ml). A putative repressor site is annotated in green. (C) DNA affinity chromatography was performed using the identified repressor site and protein enrichment values for transcription factors are plotted. Cell lysate was produced from cells grown in M9 minimal media with 0.5% glucose. Binding was performed in the presence of hypoxanthine (10 μ g/ml). Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates, and the gray shaded region represents 95% probability density region of all protein detected. (D) Identical to (B) but performed with cells containing a Δ *purR* genetic background. (E) Summary of regulatory binding sites and transcription factors that bind within the intergenic region between the genes of *yebG* and *purT*.

associated transcription factors, we next applied DNA affinity chromatography using this putative binding site sequence. In our initial attempt however, we were unable to identify any substantially enriched transcription factor (Fig. 3.10(C)). With repression observed only when cells were grown in the presence of adenine, we reasoned that the transcription factor may require a related ligand in order to bind the DNA, possibly through an allosteric mechanism. Importantly, we were able to infer an energy matrix to the putative repressor site whose sequence-specificity matched that of the well-characterized repressor, PurR ($r = 0.82$; see Fig. 3.4). We also noted ChIP-chip data of PurR that suggests it might bind within this intergenic region (Cho et al., 2011). We therefore repeated the purification in the presence of hypoxanthine, which is a purine derivative that also binds PurR (Choi and Zalkin, 1992). As shown in Fig. 3.8(C), we now observed a substantial enrichment of PurR with this putative binding site sequence. As further validation, we performed Sort-Seq once more in the adenine-rich growth condition, but in a $\Delta purR$ strain. In the absence of PurR, the putative repressor binding site disappeared (Fig. 3.8(D)), which is consistent with PurR binding at this location.

In Fig. 3.8(E) we use a "regulatory cartoon" to summarize the regulatory features between the coding genes of *purT* and *yebG*, including the new features identified by Sort-Seq. With the appearance of a simple repression architecture (Bintu et al., 2005) for the *purT* promoter, we extended our analysis by developing a thermodynamic model to describe repression by PurR. This enabled us to infer the binding energies of RNAP and PurR in absolute k_bT energies as was done in section 2.2, and we show the resulting model in Fig. 3.8(E).

The *xylE* operon is induced in the presence of xylose, mediated through binding of XylR and CRP.

The next unannotated promoter we considered was associated with expression of *xylE*, a xylose/proton symporter involved in uptake of xylose. From our analysis of the Schmidt *et al.* (Schmidt et al., 2016) data, we found that *xylE* was sensitive to xylose and proceeded by performing Sort-Seq in cells grown in this carbon source. Interestingly, the promoter exhibited essentially no expression in other media (Fig. 3.10(E)). We were able to locate the RNAP binding site between -80 bp and -40 bp relative to the *xylE* gene (Fig. 3.9(A), annotated in blue). In addition, the entire region upstream of the RNAP appeared to be involved in activating gene expression (annotated in orange in Fig. 3.9(A)), suggesting the possibility of multiple transcription factor binding sites.

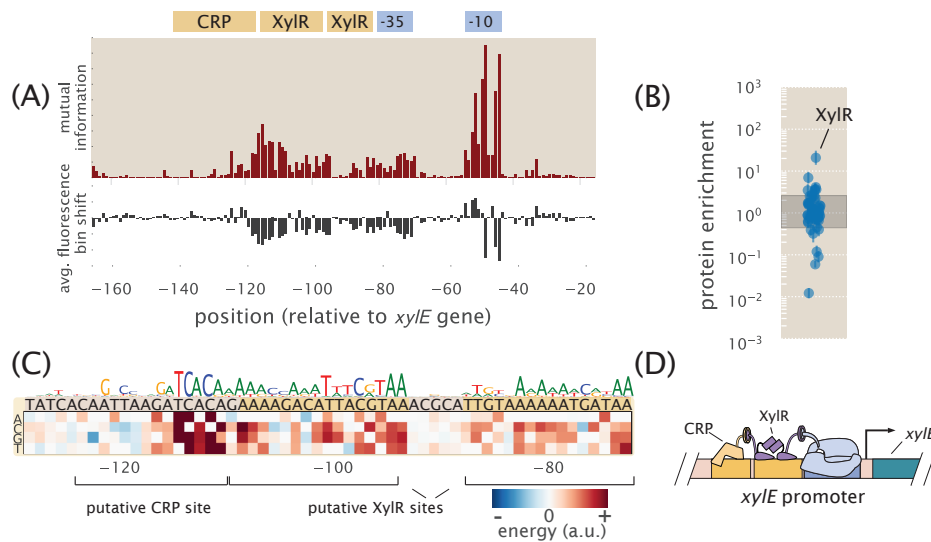


Figure 3.9: Sort-Seq identifies a set of activator binding sites that drive expression of RNAP at the *xylE* promoter. (A) Expression shifts are shown for the *xylE* promoter, with Sort-Seq performed on cells grown in M9 minimal media with 0.5% xylose. The -10 and -35 regions of an RNAP binding site (blue) and a putative activator region (orange) are annotated. (B) DNA affinity chromatography was performed using the putative activator region and protein enrichment values for transcription factors are plotted. Cell lysate was generated from cells grown in M9 minimal media with 0.5% xylose and binding was performed in the presence of xylose supplemented at the same concentration as during growth. Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates. The gray shaded region represents 95% probability density region of all proteins detected. (C) An energy matrix was inferred for the region upstream of the RNAP binding site. The associated sequence logo is shown above the matrix. Two binding sites for XylR were identified (see also Fig. 3.4) along with a CRP binding site. (D) Summary of regulatory features identified at *xylE* promoter, with the identification of an RNAP binding site and tandem binding sites for XylR and CRP.

We applied DNA affinity chromatography using a DNA target containing this entire upstream region. Due to the stringent requirement for xylose to be present for any measurable expression, xylose was supplemented in the lysate during binding with the target DNA. In Fig. 3.9(B) we plot the enrichment ratios from this purification and find XylR to be most significantly enriched. From an energy matrix inferred for the entire region upstream of the RNAP site, we were able to identify two correlated 15 bp regions (dark yellow shaded regions in Fig. 3.9(C)). Mutations of the XylR protein have been found to diminish transport of xylose (Song and Park, 1997), which in light of our result, may be due in part to a loss of activation and expression of this xylose/proton symporter. These binding sites were also similar to those found on two other promoters known to be regulated by XylR (*xylA* and *xylF* promoters), whose promoters also exhibit tandem XylR binding sites and strong binding energy predictions with our energy matrix (Fig. 3.10(F)).

Within the upstream activator region in Fig. 3.9(A) there still appeared to be a binding site unaccounted for with these tandem XylR binding sites. From the energy matrix, we were further able to identify a binding site for CRP, which is noted upstream of the XylR binding sites in Fig. 3.9(C). While we did not observe a significant enrichment of CRP in our protein purification, the most energetically favorable sequence predicted by our model, TGCGACCNAGATCACA, closely matches the CRP consensus sequence of TGTGANNNNNTCACA. In contrast to the *lac* promoter, binding by CRP here appears to depend more on the right half of the binding site sequence. CRP is known to activate promoters by multiple mechanisms (Browning and Busby, 2016), and CRP binding sites have been found adjacent to the activators XylR and AraC (Song and Park, 1997; Laikova, Mironov, and Gelfand, 2001), in line with our result. While further work will be needed to characterize the specific regulatory mechanism here, it appears that activation of RNAP is mediated by both CRP and XylR and we summarize this result in Fig. 3.9(D). The topic is considered further in Appendix A).

The *dgoRKADT* promoter is auto-repressed by DgoR, with transcription mediated by class II activation by CRP.

As a final illustration of the approach developed here, we considered the unannotated promoter of *dgoRKADT*. The operon codes for D-galactonate-catabolizing enzymes; D-galactonate is a sugar acid that has been found as a product of galactose metabolism (Cooper, 1978). We began by measuring expression from a non-mutagenized *dgoRKADT* promoter reporter to glucose, galactose, and D-

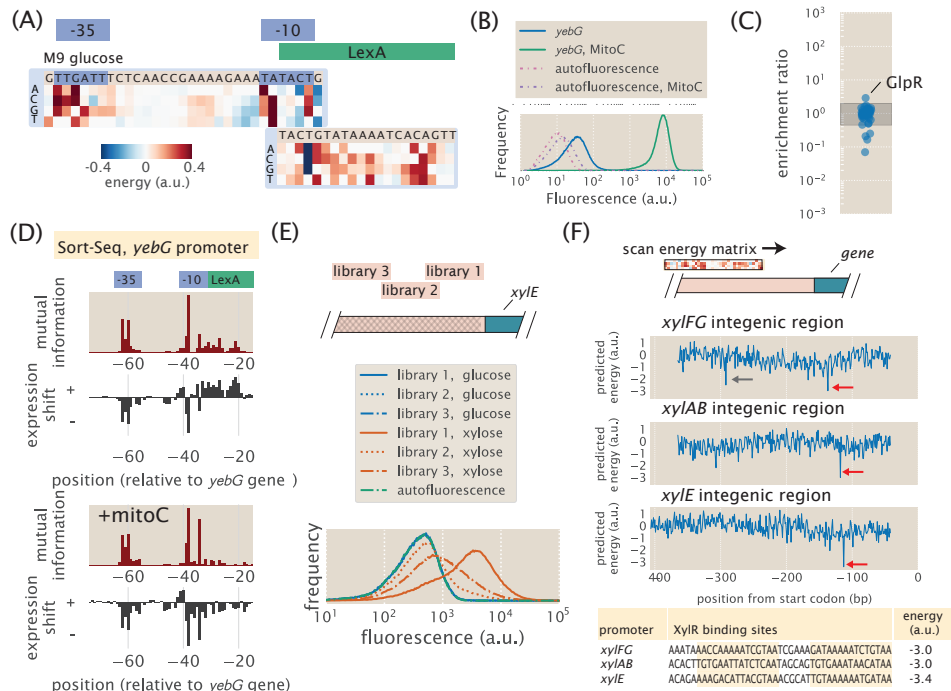


Figure 3.10: *lexA* and *yebG* regulation. (A) Energy matrices were inferred for the binding sites of LexA and RNAP. Data are from cells grown in M9 minimal media with 0.5% glucose.

(B) Fluorescence histograms for a wild-type *yebG* promoter plasmid are shown for cells grown in M9 minimal media with 0.5% glucose, and with or without mitomycin C (1 μ g/ml). Mitomycin C induces the SOS response (M. R. Lomba et al., 2006) and dramatically increases expression from the *yebG* promoter. Autofluorescence histograms refer to cells that did not contain the GFP promoter plasmid. (C) DNA affinity chromatography performed using the identified repressor site on the *purT* promoter. Cell lysate was produced from cells grown in M9 minimal media with 0.5% glucose and binding was performed in the presence of adenine (100 μ g/ml) to match the growth conditions where repression was observed. (D) Information footprints and expression shift plots are shown for the *yebG* promoter in the presence or absence of mitomycin C (1 μ g/ml). Cells were grown in M9 minimal media 0.5% glucose. (E) Fluorescence histograms are shown for the three *xylE* libraries (different mutated regions), with cells grown in M9 minimal media with either 0.5% glucose or 0.5% xylose. While xylose led to differential expression for the different libraries, cells grown in glucose were identical to autofluorescence.

galactonate. Cells grown in galactose exhibited higher expression than in glucose, as found by Schmidt *et al.* (Schmidt *et al.*, 2016) and even higher expression when cells were grown in D-galactonate (Fig. 3.10(A)). This likely reflects the physiological role provided by the genes of this promoter, which appear necessary for metabolism of D-galactonate. We therefore proceeded by performing Sort-Seq with cells grown in either glucose or D-galactonate, since these appeared to represent distinct regulatory states, with expression low in glucose and high in D-galactonate. Information footprints and expression shift plots from each growth conditions are shown in Fig. 3.11 (A). We begin by considering the results from growth in glucose (Fig. 3.11(A), top plot). Here we identified an RNAP binding site between -30 bp and -70 bp, relative to the native start codon for *dgoR* (Fig. 3.10(B)). Another distinct feature was a positive expression shift in the region between -140 bp and -110 bp, suggesting the presence of a repressor binding site. Applying DNA affinity chromatography using this target not apparent due to binding by DgoR. While only one RNAP -10 motif is clearly visible in the sequence logo shown Fig. 3.11 (C) (top sequence logo; TATAAT consensus sequence), we used simulations to demonstrate that the entire sequence logo shown can be explained by the convolution of three overlapping RNAP binding sites (See Fig. 3.10(F)).

Next we consider the D-galactonate growth condition (Fig. 3.11(A), bottom plot). Like in the expression shift plot for the Δ *dgoR* strain grown in glucose, we no longer observe the positive expression shift between -140 bp and -110 bp. This suggests that DgoR may be induced by D-galactonate or a related metabolite. However, in comparison with the expression shifts in the Δ *dgoR* strain grown in glucose, there were some notable differences in the region between -160 bp and -140 bp. Here we find evidence for another CRP binding site. The sequence logo identifies the sequence TGTGA (Fig. 3.11(D), bottom logo), which matches the left side of the CRP consensus sequence. In contrast to the *lac* and *xylE* promoters however, the right half of the binding site directly overlaps with where we would expect to find a -35 RNAP binding site. This type of interaction by CRP has been previously observed and is defined as class II CRP dependent activation (Browning and Busby, 2016), though this sequence-specificity has not been previously described.

In order to isolate and better identify this putative CRP binding site we repeated Sort-Seq in *E. coli* strain JK10, grown in 500 μ M cAMP. Strain JK10 lacks adenylate cyclase (*cyaA*) and phosphodiesterase (*cpdA*), which are needed for cAMP synthesis and degradation, respectively, and is thus unable to control intracellular cAMP levels

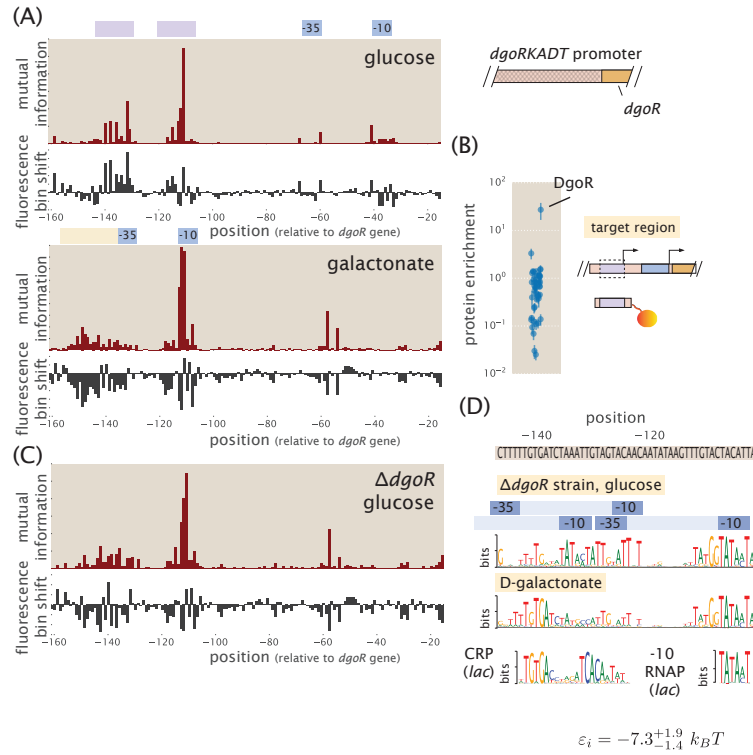


Figure 3.11: The *dgoRKADT* promoter is induced in the presence of D-galactonate due to loss of repression by DgoR and activation by CRP. (A) Expression shifts due to mutating the *dgoRKADT* promoter are shown for cells grown in M9 minimal media with either 0.5% glucose (top) or 0.23% D-galactonate (bottom). Regions identified as RNAP binding sites (-10 and -35) are shown in blue and putative activator and repressor binding sites are shown in yellow and purple, respectively. (B) DNA affinity purification was performed targeting the region between -145 to -110 of the *dgoRKADT* promoter. The transcription factor DgoR was found most enriched among the transcription factors plotted. Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates, and the gray shaded region represents 95% probability density region of all proteins detected. (C) Sequence logos were inferred for the most upstream 60 bp region associated with the upstream RNAP binding site annotated in (A). Multiple RNAP binding sites were identified using Sort-Seq data performed in a $\Delta dgoR$ strain, grown in M9 minimal media with 0.5% glucose. Below this, a sequence logo was also inferred using data from Sort-Seq performed on wild-type cells, grown in D-galactonate, identifying a CRP binding site (class II activation; Browning and Busby, 2004). (D) Summary of regulatory features, with sequence logos, for features identified at *dgoRKADT* promoter, with the identification of multiple RNAP binding sites, and binding sites for DgoR and CRP. An initial estimate of $-7.3 k_B T$ was determined for the interaction energy between CRP and RNAP, ε_i .

necessary for activation by CRP (derivative of TK310 (Kuhlman, Z. Zhang, et al., 2007)). Growth in the presence of 500 μM cAMP provided strong induction from the *dgoRKADT* promoter and resulted in a sequence logo at the putative CRP binding site that even more clearly resembled binding by CRP (Fig. 3.10(E)). This is likely because expression is now dominated by the CRP activated RNAP binding site. Importantly, this data allowed us to further infer the interaction energy between CRP and RNAP, which we estimate to be $-7.3 k_bT$ (further detailed Appendix A). We summarize the identified regulatory features in Fig. 3.11(D).

3.3 Discussion

We have established a systematic and scalable procedure for dissecting the functional mechanisms of previously uncharacterized regulatory sequences in bacteria. A massively parallel reporter assay, Sort-Seq (Kinney and Callan, 2010), is used to first elucidate the locations of functional transcription factor binding sites. DNA oligonucleotides containing these binding sites are then used to enrich the cognate transcription factors and identify them by mass spectrometry analysis. Information-based modeling and inference of energy matrices that describe the DNA binding specificity of regulatory factors provide additional insight into transcription factor identity and the growth condition dependent regulatory architectures.

To validate this approach we examined four previously annotated promoters. Our Sort-Seq results were in good agreement with established knowledge for *lacZYA*, *relBE*, *marRAB* (Oehler et al., 1990; Kinney and Callan, 2010; Garcia and Phillips, 2011; Bech et al., 1985; Gotfredsen and Gerdes, 1998; Overgaard, Borch, Jørgensen, et al., 2008; Seoane and Levy, 1995; Alekshun and Levy, 1997). For the *yebG* promoter, our approach corrected an error in a previous annotation. DNA affinity chromatography experiments on these promoters were found to be highly sensitive. In particular, LacI was unambiguously identified with the weak O3 binding site, even though LacI is present in only about 10 copies per cell (Garcia and Phillips, 2011).

Emboldened by this success, we then studied promoters having little or no prior regulatory annotation: *purT*, *xylE*, *dgoR*. Through extensive modeling of the Sort-Seq data and DNA affinity chromatography of many identified binding sites, our analysis led to a collection of new regulatory hypotheses. For the *purT* promoter, we identified a simple repression architecture (Bintu et al., 2005), with repression by PurR. The *xylE* promoter was found to undergo activation only when cells are

grown in xylose, likely due to allosteric interaction between the activator XylR and xylose, and activation by CRP (Song and Park, 1997; Laikova, Mironov, and Gelfand, 2001). Finally, in the case of *dgoR*, the base pair resolution allowed us to tease apart multiple overlapping binding sites. In particular, we were able to identify multiple RNAP binding sites along the length of the promoter. Of these, one set of RNAP binding sites were repressed by DgoR when cells were grown in glucose, but activated through class II activation by CRP when D-galactonate was used as the sole carbon source. We view these results as a critical first step in the quantitative dissection of transcriptional regulation, which will ultimately be needed for a predictive understanding of how such regulation works. The regulatory cartoons shown in Fig. 3.8(D) and Fig. 3.9(D) will serve as a starting point for further mathematical dissection of these promoters and will lead to a series of quantitative predictions for how the different promoters work.

There are a number of ways to further increase the resolution and throughput of the methods we have described. Microarray-synthesized promoter libraries allow multiple loci to be studied simultaneously as we prove in Chapter 4. Landing pad technologies for chromosomal integration (Kuhlman and E. C. Cox, 2010; H. Zhang et al., 2016), should enable massively parallel reporter assays to be performed in chromosomes instead of on plasmids. Techniques that combine these assays with transcription start site readout (Vvedenskaya, Goldman, and Nickels, 2015) may further allow the molecular regulators of overlapping RNAP binding sites to be deconvolved, or the contributions from separate RNAP binding sites, like those observed on the *dgoR* promoter, to be better distinguished.

Although our work was directed toward regulatory regions of *E. coli*, there are no intrinsic limitations that restrict the analysis to this organism. Rather, it should be applicable to any bacterium that supports efficient transformation by plasmids. And although we have focused on bacteria, our general approach should be feasible in a number of eukaryotic systems – including human cell culture – using massively parallel reporter assays (Melnikov et al., 2012; Kheradpour et al., 2013; Patwardhan et al., 2012) and DNA-mediated protein pull-down methods (Mittler, Butter, and M. Mann, 2009; Mirzaei et al., 2013) that have already been established.

3.4 Methods

Our intention was to construct a systematic and scalable experimental pipeline that would be applicable to the general objective of discovering regulatory architectures

in generic bacteria. In this section we describe the work flow required by this pipeline with the aim of giving a sense of how each of the steps is carried out. We begin with a description of how we construct the mutated promoters used in the Sort-Seq experiment itself. Next we describe how the fluorescence levels are measured in a FACS machine and how the sorting and sequencing are performed. After that, the remainder of the Methods section focuses on the steps required to perform DNA affinity chromatography and mass spectrometry, which is necessary to identify the transcription factors that bind to the putative binding sites identified in the Sort-Seq procedure.

Sort-Seq libraries

Mutagenized single-stranded oligonucleotide pools were purchased from Integrated DNA Technologies (Coralville, IA), with a target mutation rate of 9%. In the case of the *lacZ* promoter, the associated Sort-Seq data was also used in the analysis in (Razo-Mejia et al., 2014). The mutation rate for this library was approximately 3%. Each oligonucleotide was PCR amplified in order to produce double-stranded inserts, which were inserted into a PCR amplified plasmid backbone (i.e. vector) of pJK14 (Kinney and Callan, 2010) by Gibson Assembly (Gibson et al., 2009) (New England Biolabs, MA, USA). Note however that in the construction of the *lacZ* promoter, assembly was performed using restriction cloning as in Kinney and Callan, 2010. The template plasmid used for amplification of the backbone contained the toxic gene *ccdB* in place where the library was to be inserted. In this way any bacteria that took up any of the initial plasmid used in the PCR amplification would be removed from the population via negative selection due to toxicity by the *ccdB* gene (propagated in the immune strain DB3.1). This helped ensure that no template plasmid was propagated into the final plasmid library (see methods in reference (Kinney and Callan, 2010) for more detail). The plasmid is maintained at low copy numbers (about 5 copies per cell) by the SC101 origin of replication (Lutz, 1997).

For each library construction, 40 ng of insert and 50 ng of backbone were combined in a 20 μ L Gibson assembly reaction. To achieve high transformation efficiency, reaction buffer components from the Gibson Assembly reaction were removed by drop dialysis and cells were transformed by electroporation of freshly prepared cells. Following an initial outgrowth in 1 mL of SOC media, cells were diluted into 50 mL of LB media and grown overnight under kanamycin selection. Transformation typically yielded $10^6 - 10^7$ colonies as assessed by plating 100 μ L of cells diluted

1 : 10⁴ onto an LB plate containing kanamycin.

3.5 Bacterial strains

All *E. coli* strains used in this work were derived from K12 MG1655, with deletion strains generated by the lambda red recombinase method (Datsenko and Wanner, 2000; Sawitzke et al., 2007). In the case of deletions for *lysA* ($\Delta lysA :: kan$), *purR* ($\Delta purR :: kan$), and *xylE* ($\Delta xylE :: kan$), strains were obtained from the Coli Genetic Stock Center (CGSC, Yale University, CT, USA) and transferred into a fresh MG1655 strain using P1 transduction (Thomason, Costantino, and Court, 2007). The others were generated in house and include the following deletion strains: $\Delta lacIZYA$, $\Delta relBE :: kan$, $\Delta marRAB :: kan$, $\Delta marR :: kan$, $\Delta dgoR :: kan$.

Here we describe the approach used to generate these deletion strains. Briefly, an overnight culture of MG1655 containing the plasmid pSIM6 was diluted 1:100 in 50 mL LB media and grown to an OD₆₀₀ of ≈ 0.4 at 30°C. The culture was immediately placed in a water bath shaker at 43°C for 15 minutes and then cooled in an ice bath for 10 minutes. Cells were then spun down for 10 minutes (4,000 g, 4°C) and resuspended on ice in 50 mL of chilled water. This was repeated three times before resuspending in 200 μ L of chilled water to generate competent cells. Homologous primer extension sequences for the appropriate gene were obtained from Baba et al., 2006 and used to generate linear DNA containing a kanamycin resistance gene insert by PCR, which contained homology for the region on the chromosome to be deleted (Datsenko and Wanner, 2000). Electroporation of the competent cells was performed using 1 μ L purified PCR product (about 100 ng DNA), mixed with 50 μ L cells. Cells were immediately resuspended in 750 μ L SOC media and placed on a shaker at 30°C for outgrowth, for 90-120 minutes. Cells were then plated on an LB-agar plate containing kanamycin (30 μ g/mL) and grown overnight at 30°C. The deletions were confirmed by both colony PCR and sequencing. After confirmation, the deletion was transferred to a clean MG1655 strain through P1 transduction and selection on kanamycin. In the case of the lysine auxotrophic strain, we also confirmed deletion of *lysA* by checking that the cells were unable to grow in M9 minimal media unless lysine was supplemented (40 μ g/mL).

To generate strains with different LacI tetramer copy numbers per cell (associated with data in Supplemental Fig. 3.6(C)), the LacI constructs from Garcia and Phillips, 2011 were P1 transduced into the $\Delta lacIZYA$ strain (integrated at the *ybcN* locus).

Sort-Seq fluorescence sorting

Cells were grown to saturation in LB and then diluted 1:10,000 into the appropriate growth media for the promoter under consideration. For cells grown in 0.23% D-galactonate in M9 minimal media, D-galactonate appeared to form precipitates, but cells otherwise appeared to grow normally. Upon reaching an OD₆₀₀ of about 0.3, the cells were washed two times with chilled PBS by spinning down the cells at 4000 rpm for 10 minutes at 4°C. After washing with PBS, they were then diluted two fold with PBS to an OD of 0.1-0.15. This diluted cell solution was then passed through a 40 μ m cell strainer to eliminate large clumps of cells.

A Beckman Coulter MoFlo XDP cell sorter was used to obtain fluorescence histograms of between 200,000 and 500,000 cell events per culture. For libraries, these histograms were used to set the four binning gates, which each covered 15% of the histogram. During sorting of each library, 500,000 cells were collected into each of the four bins. Finally, sorted cells were re-grown overnight in 10 mL of LB media, under kanamycin selection.

Sort-Seq sequencing

The contents of each bin were miniprepmed following overnight growth (Qiagen, Germany). PCR was used to amplify the mutated region from each plasmid for Illumina sequencing. The primers contained Illumina adapter sequences as well as barcode sequences that enabled pooling of the samples. Sequencing was performed by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech (HiSeq 2500) or NGX Bio (NextSeq sequencer; San Francisco, CA). Single-end 100 bp or paired-end 150 bp flow cells were used, with a target read count of about 500,000 sequences per library bin. Joining of paired-end reads was performed with the FLASH tool (Magoc and Salzberg, 2011). For quality filtering, we collected sequences whose barcodes had a PHRED score greater than 20 at each position. Some libraries also contained non-mutagenized regions, and upon checking these, sequences that did not contain the expected sequence were excluded from our analysis. The total number of useful reads available to produce information footprints, fluorescence bin shift plots, energy weight matrices, and sequence logos from each Sort-Seq experiment generally ranged between 300,000 to 2,000,000 reads. Energy matrices were inferred using Bayesian parameter estimation with an error-model-averaged likelihood as previously described (Kinney and Callan, 2010; Kinney and Atwal, 2014), using the MPATHic software (Ireland and Kinney, 2016). A more detailed description of the data analysis procedures is available in Appendix A.

Lysate preparation and SILAC incorporation

SILAC labeling (Ong et al., 2002) was implemented by growing cells in either the stable isotopic form of lysine ($^{13}\text{C}_6\text{H}_{14}^{15}\text{N}_2\text{O}_2$), referred to as the heavy label, or natural lysine, referred to as the light label. By differentially labeling cell lysates we were able to simultaneously quantify the abundance of protein between two DNA affinity purification samples (i.e. one using a target binding site sequence and another as a reference control). This allows us to identify whether any protein shows a preference for the target binding site sequence.

To confirm heavy lysine was being incorporated, MG1655 $\Delta\text{lysA}::\text{kan}$ cells from an overnight M9 minimal media culture were diluted 1:200 and 1:1,000, and grown in 1 mL M9 minimal media supplemented with $\mu\text{g/mL}$ heavy lysine. Following approximately 7 and 10 cell divisions, cells were resuspended in lysis buffer (50 mM HEPES pH 7.5, 70 mM potassium acetate, 5mM magnesium acetate 0.2% (w/v) n-dodecyl-beta-D-maltoside, Roche protease inhibitor cOmplete tablet) and lysed by performing 10 freeze-thaw cycles with dry ice. Cellular debris were removed by centrifugation at 14000 g at 4°C on a tabletop centrifuge. Finally cellular lysates were prepared for mass spectrometry by in-solution digestion with endoproteinase Lys-C (Promega, Madison, WI). Digestion was performed as described elsewhere (Wiśniewski et al., 2009) and labeling of the heavy isotope was confirmed by mass spectrometry measurement. In addition, we also characterized the SILAC enrichment ratio measurement by directly combining measurements from heavy and light lysates over a range from 0.1:1 to 1,000:1 heavy:light (see Supplemental Section 3.12).

To generate each lysate for DNA affinity purification experiments, an overnight starter culture of cells was grown in LB media supplemented with kanamycin (30 $\mu\text{g/mL}$). An aliquot was washed twice in M9 minimal media and resuspended to an OD600 of ≈ 1.0 . For both heavy and light labeling, 500 mL M9 minimal media was then inoculated at 1:5,000 and grown to an OD600 of ≈ 0.6 (supplemented with the appropriate lysine; 40 $\mu\text{g/mL}$). Cultures were pelleted using an ultracentrifuge (8,000 g, 40 minutes) at 4°C and resuspended in chilled 20 ml lysis buffer containing 1 % (w/v) n-dodecyl-beta-maltoside. The pellets could also be stored at -80°C for later use. Cells were then lysed with a Cell Disruptor (CF Range, Constant Systems Ltd., UK) and following removal of debris by centrifugation, concentrated to $\sim 150 \text{ mg/ml}$ using Amicon Ultra-15 centrifugation units (3kDaMWCO, Millipore). This provided about 600 μL of lysate, suitable for about six 80 μL DNA affinity

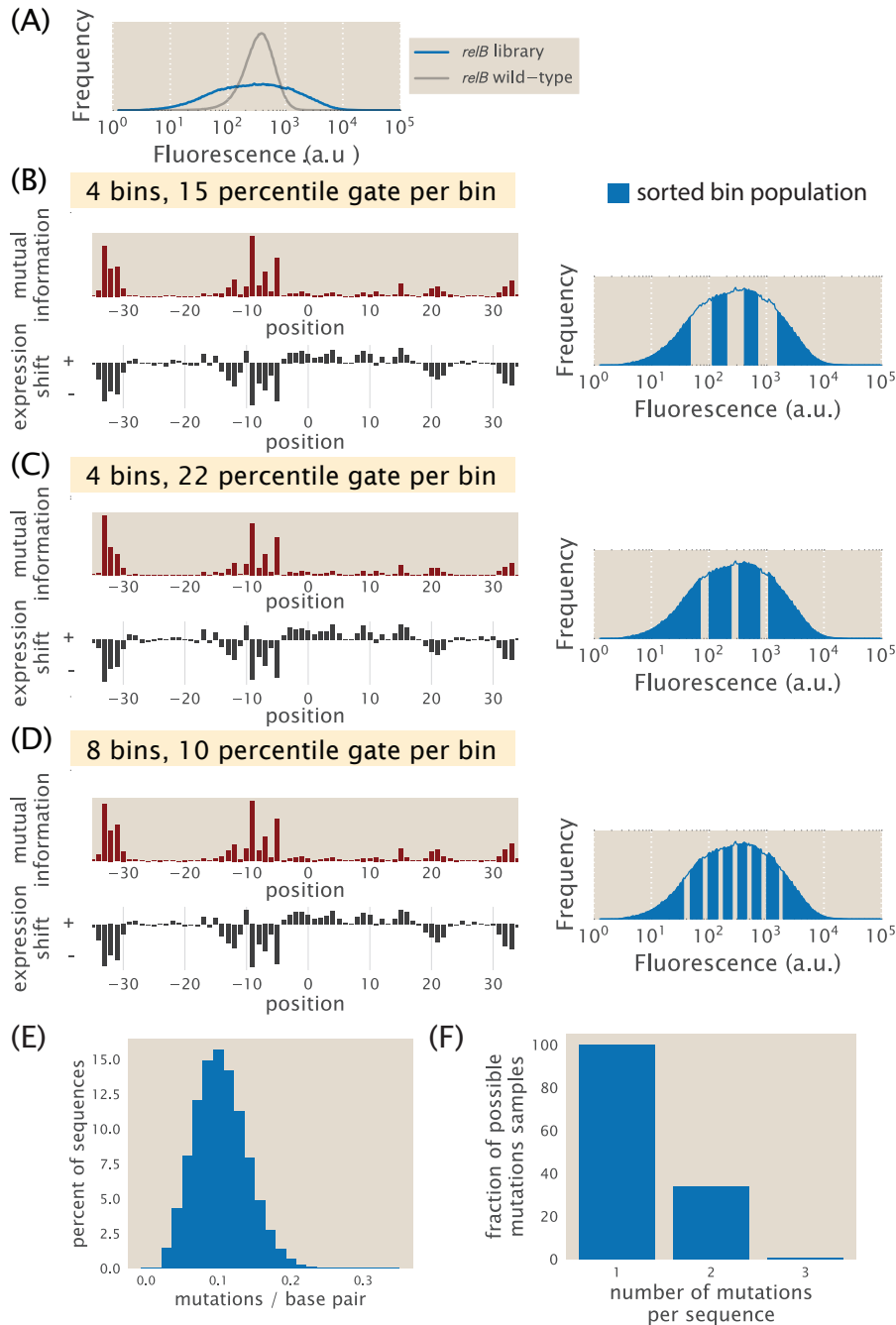


Figure 3.12: Related to Fig. 3.2 and Fig. 3.3. Analysis of the library mutation spectrum and effect of Sort-Seq sorting conditions. (A) Here we used our *relBE* promoter library to test whether the sorting procedure influenced our Sort-Seq data analysis. The fluorescence histogram of the wild-type promoter plasmid (single clonal population) and the mutated library for the *relB* promoter are shown. Expression shifts and information footprints are shown for cells sorted under three different scenarios in (B)-(D). In (B) cells were sorted using the approach of the main text where cells were sorted into 4 bins, each containing 15% of the population.

purifications. Total protein concentrations was assayed using the Bradford reagent (Sigma-Aldrich, St. Louis, MO). Following adjustment of protein concentration, sheared salmon sperm competitor DNA was added to the lysates (1 $\mu\text{g}/\text{mL}$; Life Technologies, Carlsbad, CA) and incubated for 10 minutes at 4°C. Finally, following centrifugation at 14,000 g to remove insoluble matter, lysates were either placed on ice or stored at 4°C prior to use.

Preparation of DNA-tethered magnetic beads

DNA affinity chromatography was performed by incubating cell lysate with magnetic beads (Dyanbeads MyOne T1, Life Technologies, Carlsbad) containing tethered DNA. The DNA was tethered through a linkage between streptavidin on the beads and biotin on the DNA. Note single-stranded DNA was purchased from Integrated DNA Technologies with the biotin modification on the 5' end of the oligonucleotide sense strand. Briefly, DNA was suspended in annealing buffer (20 mM Tris-HCl, 10 mM MgCl₂, 100 mM KCl) to 50 μM . Complementary strands were annealed by mixing 30 μL of the sense strand and 40 μL of the complement strand. Excess complement strand ensured all biotinylated-DNA would be in a double stranded form. Annealing was then performed using a thermocycler: 90°C for 5 minutes, gradient from 90°C to 65°C at 0.1 °C/sec, incubated from 10 minutes at 65°C and allowed to return to room temperature on the thermocycler. Prior to attaching DNA, 150 μL beads were washed twice with 600 μL TE buffer (10 mM Tris-HCL pH 8.0, 1 mM EDTA) and then twice with DW buffer (20 mM Tris-HCL pH 8.0, 2 M NaCl, 0.5 mM EDTA (Mittler, Butter, and M. Mann, 2009)). Approximately 640 pmol of DNA were then diluted to 600 μL in DW Buffer and incubated with the washed beads overnight at 4°C and on a rotatory wheel. Bound DNA was measured by determining the DNA concentration before and after incubation with beads using a NanoDrop (Thermo Scientific, Waltham, MA). Finally, beads were washed once with 600 μL TE buffer and three washes of 600 μL DW buffer, and resuspended in 150 μL DW buffer.

DNA affinity chromatography

Prior to DNA affinity purification the DNA tethered beads were incubated with blocking buffer (20 mM Hepes, pH 7.9, 0.05 mg/ml BSA, 0.05 mg/ml glycogen, 0.3 M KCl, 2.5 mM DTT, 5 mg/ml polyvinylpyrrolidone, 0.02% (w/v) n-dodecyl- β -D-maltoside; about 1.3 mL/mg beads (Mittler, Butter, and M. Mann, 2009) for one hour at 4°C for passivation. Excess blocking buffer was removed by washing

the beads twice with 600 μ L lysis buffer. The cell lysates were also incubated with washed magnetic beads that contained no tethered DNA. Following removal of these beads, cell lysates were incubated on a rotating wheel with the DNA tethered beads for approximately five hours at 4°C. Beads were then recovered with a magnet and washed three times using an equivalent volume of lysis buffer. The beads were then washed once more, but with NEB Buffer 3.1 (New England Biolabs, MA, USA). Both purifications (with the target DNA and reference control) were then combined by re-suspending in 50 μ L NEB Buffer 3.1. To this suspension, 10 μ L of the restriction enzyme PstI (100,000 units/mL, New England Biolabs) was added and incubated for 1.5 hours at 25°C. PstI cleaves the sequence CTGCAG, which was included between the biotin label and binding site sequence, allowing the DNA to be released from the magnetic beads. The beads were then removed and the samples diluted with 4x SDS-PAGE sample buffer. After incubation for five minutes at 95°C, the samples were then loaded on a SDS-PAGE gel (Any k_D Mini-PROTEAN TGX Precast Protein Gels, 10-well, 50 μ L; BioRad, CA, USA) and gel electrophoresis was performed for 45-55 minutes (200V) to separate proteins by size. The gel was stained using the Colloidal Blue Staining Kit (ThermoFisher Scientific, MA, USA) for visualization. Note that in general, we purified proteins from a heavy lysate using DNA containing the target binding site sequence, while devoting the light lysate to a control DNA sequence. However, for our LacI and RelBE, we also performed the alternative scenario (i.e. target binding site sequence purified with the light lysate). We did not observe major differences between either approach and therefore continued in our other experiments by purifying with the target binding site sequence in the heavy lysate.

In-gel digestion for mass spectrometry

After destaining, the gel was cut into four sections, each of which was cut into small pieces for in-gel digestion. The gel pieces were reduced, alkylated, and digested by endoproteinase Lys-C overnight at 37°C. This enzymatically cleaves proteins after lysine residues and is necessary for determining whether detected peptides are from the light or heavy lysine labeled purification. Digested peptides were extracted from gel and lyophilized. The peptide samples were further purified using StageTips to remove residual salts (Rappsilber, Mann, and Y. Ishihama, 2007). The extracts were re-suspended in 0.2% formic acid.

LC-MS/MS analysis and protein quantification

Liquid chromatography-tandem-mass spectrometry (LC-MS/MS) experiments were carried out as previously described (Kalli and Hess, 2012). The LacI target purification experiments were performed on a nanoflow LC system, EASY-nLC II coupled to a hybrid linear ion trap Orbitrap Classic mass spectrometer equipped with a Nanospray Flex Ion Source (Thermo Fisher Scientific). The in-gel digested peptides were directly loaded at a flow rate of 500 nL/min onto a 16-cm analytical HPLC column (75 μ m ID) packed in-house with ReproSil-Pur C18AQ 3 μ m resin (120 Å pore size, Dr. Maisch, Ammerbuch, Germany). The column was enclosed in a column heater operating at 45°C. After 30 min of loading time, the peptides were separated in a solvent gradient at a flow rate of 350 nL/min. The gradient was as follows: 0-30 % B (80 min), and 100% B (10 min). The solvent A consisted of 97.8 % H_2O , 2% ACN, and 0.2% formic acid and solvent B consisted of 19.8% H_2O , 80 % ACN, and 0.2 % formic acid. The Orbitrap was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z =400–1600) in the Orbitrap (resolution 100,000) and subsequent 15 CID MS/MS scans (Top 15 method) in the linear ion trap. Collision induced dissociation (CID) was performed at normalized collision energy of 35 % and 30 msec of activation time. All other measurements were performed on a hybrid ion trap-Orbitrap Elite mass spectrometer (Thermo Fisher Scientific), which provided greater detection sensitivity and other fragmentation techniques as described. The Orbitrap was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z =400–1,800) in the Orbitrap (resolution 120,000) and subsequent 5 MS/MS scans also acquired in Orbitrap with 15,000 resolution. The MS/MS spectra were acquired for the top 5 ions alternating between higher collision dissociation (HCD) and electron transfer dissociation (ETD) fragmentations that are well suited for higher charge peptides. Higher collision dissociation was performed at a normalized collision energy of 30 % and electron transfer dissociation reaction time was set to 100 msec. The analytical column for this instrument was a PicoFrit column (New Objective, Woburn, MA) packed in house with ReproSil-Pur C18AQ 1.9 μ m resin (120Å pore size, Dr. Maisch, Ammerbuch, Germany) and the column was heated to 60°C. The peptides were separated either with a 90 or 60 min gradient (0-30% B in 90 min or 0-30% B in 60 min) at a flow rate of 220 nL/min.

Thermo RAW files were processed using MaxQuant (v. 1.5.3.30) (Jürgen Cox and Mann, 2008; Jürgen Cox et al., 2009). Spectra were searched against the UniProt E. coli K12 database (4318 sequences) as well as a contaminant database

(256 sequences). Precursor ion mass tolerance was 4.5 ppm after recalibration by MaxQuant. Fragment ion mass tolerance was 20 ppm for high-resolution HCD and ETD spectra, and 0.5 Da for low-resolution CID spectra. Variable modifications included oxidation of methionine and protein N-terminal acetylation. Carboxyamidomethylation of cysteine was specified as a fixed modification. LysC was specified as the digestion enzyme and up to two missed cleavages were allowed. A decoy database was generated by MaxQuant and used to set a score threshold so that the false discovery rate was less than 1 % at both the peptide and protein level. For all experiments match between runs and re-quantify were enabled. One evidence ratio per replicate per protein was required for quantitation. To calculate the overall protein ratio, the un-normalized protein replicate ratios were log transformed and then shifted so that the median protein log ratio within each replicate was zero (i.e., the median protein ratio was 1:1). The overall experimental log ratio was then calculated from the average of the replicate ratios.

Data analysis, code, and data curation

Additional details about the data analysis and characterization of Sort-Seq and DNA affinity chromatography can be found in the Supplemental material. The identification of regulated operons shown in Fig. 3.1 was performed using the annotated operons listed on RegulonDB (Gama-Castro et al., 2016), which are based on manually curated experimental and computational data. An operon was considered to be regulated if it had at least one transcription factor binding site associated with it. All code used for processing data and plotting, as well as the final processed data can be found on our GitHub repository (https://github.com/RPGroup-PBoC/Sort-seq_belliveau). Thermo RAW files for mass spectrometry are available on the jPOSTrepo repository (Okuda et al., 2017) under accession code PXD007892. Sort-Seq sequencing files are available on the Sequence Read Archive (accession code SRP121362).

Acknowledgements

We thank David Tirrell, Bradley Silverman, and Seth Lieblisch for access and training for use of their Beckman Coulter MoFlo XDP cell sorter. We thank Jost Vielmetter and Nina Budaeva for access and training for use on their Cell Disruptor. We also thank Hernan Garcia, Manuel Razo-Mejia, Griffin Chure, Suzannah Beeler, Heun Jin Lee, Justin Bois, and Soichi Hirokawa for useful advice and discussion. Hernan Garcia also helped generate Figure 1. This work was supported by La

Fondation Pierre-Gilles de Gennes, the Rosen Center at Caltech, and the National Institutes of Health DP1 OD000217 (Director's Pioneer Award), R01 GM085286, and 1R35 GM118043-01 (MIRA), the Gordon and Betty Moore Foundation through GBMF227, the National Institutes of Health 1S10RR029594-01A1 and the Beckman Institute. NB is a Howard Hughes Medical Institute International Student Research fellow.

3.6 Supplemental Information: Characterization of library diversity and sorting sensitivity.

Sort-Seq of the *rel* promoter using different sorting conditions.

In the work of the main text, Sort-Seq was performed by sorting cell libraries into four bins based on their fluorescence, each containing about 15 percent of the population. The remaining population was not collected and was discarded to waste. Due to the variability in expression of a single clonal population (Fig. 3.12(A)), sorting into a larger number of narrower bins was not expected to provide significant additional resolution for the sequence-dependent fluorescence distribution. Given the success in identifying the known regulatory binding sites of the *lacZ*, *relB*, and *marR* promoters, and agreement between the inferred sequences logos and available sequence logos (see Supplemental Fig. 3.4), these conditions appeared to provide sufficient information to accurately analyze our libraries.

However, in order to further confirm that our results were not being influenced by the specific sorting scheme, we also tested several other sorting conditions using our *relB* promoter library. Here cells were sorted into either 4 or 8 bins, with a sorting gate containing between 10 and 22 percent of the population per bin. The associated expression shift plots and information footprints (defined in Supplemental Section are shown in Fig. 3.12(B)-(D). In general we found little difference between each of these experiments. Energy matrices for the binding sites were similarly in agreement, with a Pearson correlation coefficient between matrix parameters generally greater than 0.9 across the different conditions tested.

3.7 Analysis of library diversity using data from the *mar* promoter.

Here we provide additional characterization of the mutagenized promoter libraries, using a library from the *marR* promoter as a representative example (70 bp region containing RNAP and MarR repressor sites). With the exception of the *lacZ* promoter, all library oligonucleotide pools were purchased from Integrated DNA Technologies (USA) with a target mutation rate of nine percent per nucleotide po-

sition. For the *lacZ* promoter library, we purchased an oligonucleotide pool using their Ultramer branded technology to allow for a longer mutagenized region that covered the known set of regulatory binding sites. While we intended to have a similar mutation rate, we found a mutation rate closer to three percent per nucleotide position. While unexpected, this allowed us to test two different mutation rates in our initial validation of the methodology using well-characterized promoters.

To get a better sense of how the mutation rate varies across the libraries, we plot a histogram of the number of mutations per base pair for the entire set of sequences found in the *marR* promoter library (Fig. 3.12(E)). While we obtained an average mutation rate of 10.4% in this library, close to our target rate of 9%, there is some variability in this mutation rate as might be expected given that the incorporation of mutations in the DNA synthesis procedure is a random process. Since we are using these sequence data sets to infer sequence-specific models of binding between DNA and transcription factors, it was also of interest to consider the mutational coverage found within the library. As shown in Fig. 3.12(F), all single-point mutations and a large fraction of two-point mutations were present within the library. Due to the large number of possible three point mutants in a 60 bp region, only a small subset of the possible sequences will be found in the library.

3.8 Supplemental Information: Generation of sequence logos.

Sequence logos provide a simple way to visualize the sequence specificity of a transcription factor to DNA, as well as the amount of information present at each position (Schneider and Stephens, 1990). Here we describe how we generate them using either known genomic binding sites or the energy matrices that were determined from our Sort-Seq data. In each case we need to calculate a $4 \times L$ position weight matrix for a binding site of length L , which is used to estimate the position-dependent information content needed to construct a sequence logo.

3.9 Generating position weight matrices from known genomic binding sites.

From RegulonDB, we find there are $N_g = 260$ known binding sites for CRP on the *E. coli* genome (Gama-Castro et al., 2016). To construct a position weight matrix using these genomic binding sites, we must first align all the sequences and determine the nucleotide statistics at each position. Specifically, we count the number of each nucleotide, N_{ij} , at each position along the binding site. Here the subscript i refers to the position, while j refers to the nucleotide, A , C , G , or T . We can then calculate a position probability matrix (also $4 \times L$) where each entry is found by dividing these

counts by the total number of sequences in our alignment,

$$p_{ij} = \frac{N_{ij}}{N_g}. \quad (3.1)$$

Note that in situations where the number of aligned sequences is small (e.g., less than five), pseudocounts (Nishida, Frith, and Nakai, 2009) are often added to regularize the probabilities of the counts in the calculation of position probabilities,

$$p_{ij} = \frac{N_{ij} + B_p}{N_g + 4B_p}, \quad (3.2)$$

where B_p is the value of the pseudocount. The argument for their use is that when selecting from a small number of binding site sequences, just by chance infrequent nucleotides will be absent, and assigning them a probability (p_{ij} , noted above) of zero may be too stringent of a penalty (Nishida, Frith, and Nakai, 2009). We let $B_p = 0.1$. In the limit of zero binding site sequences (i.e., with no sequences observed), this will result in probabilities p_{ij} approximately equal to the background probability used in calculating the position weight matrix below (and a non-informative sequence logo).

Finally, the values of the position weight matrix are found by calculating the log probabilities relative to a background model (Stormo, 2000).

$$\text{PWM}_{ij} = \log_2 \frac{p_{ij}}{b_j}. \quad (3.3)$$

The background model reflects assumptions about the genomic background of the system under investigation. For instance, in many cases it may be reasonable to assume each base is equally likely to occur. Given that we know the base frequencies for *E. coli*, The background model reflects assumptions about the genomic background of the system under investigation. For instance, in many cases it may be reasonable to assume each base is equally likely to occur. Given that we know the base frequencies are $A = 0.246$, $C = 0.254$, $G = 0.254$, $T = 0.246$ for strain MG1655 (BioNumbers ID 100528, <http://bionumbers.hms.harvard.edu>). From Eq. 3.3 we can see that the value at the i^{th} ; j^{th} position will be zero if the probability, p_{ij} , matches that of the background model, but non-zero otherwise. This reflects the fact that base frequencies matching the background model tell us nothing about the binding preferences of the transcription factor, while deviation from this background frequency indicates sequence specificity.

3.10 Generating position weight matrices from Sort-Seq data.

Next we construct a position weight matrix using the CRP energy matrix from our Sort-Seq data. Here we appeal to the result from Berg and von Hippel, that the logarithms of the base frequencies above should be proportional to their binding energy contributions (Berg and Hippel, 1987; Stormo, 2000). Berg and von Hippel considered a statistical mechanical system containing L independent binding site positions, with the choice of nucleotide b_j at each position corresponding to a change in the energy level by ε_{ij} relative to the lowest energy state at that position. ε_{ij} corresponds to the energy entry in our energy matrix, scaled to absolute units, $A\theta_{ij}+B$, where θ_{ij} is the i^{th}, j^{th} entry. An important assumption is that all nucleotide sequences that provide an equivalent binding energy must have equal probability of being present as a binding site. In this way, we can relate the binding energies considered here to the statistical distribution of binding sites in the previous section. The probability p_{ij} of choosing nucleotide b_j at position i for protein binding will then be proportional to probability that position i has energy ε_{ij} . Specifically, the probabilities will be given by their Boltzmann factors normalized by the sum of states for all nucleotides,

$$p_{ij} = \frac{b_j e^{-\beta A \theta_{ij}}}{\sum_{j=A}^T b_j e^{-\beta A \theta_{ij}}}, \quad (3.4)$$

where $\beta = \frac{1}{k_b T}$, where k_b is the Boltzmann constant and T is the temperature in Kelvin.

One difficulty that arises when we use energy matrices that are not in absolute energy units is that we are left with an unknown scale factor A , preventing calculation of p_{ij} . We appeal to the expectation that mismatches usually involve an energy cost of 1-3 $k_b T$ (Lässig, 2007). In other work within our group, we have found this to be a reasonable assumption for LacI. Therefore, we approximate it such that the average cost of a mutation $\langle A \times \theta_{ij} \rangle = 2k_b T$. We can then calculate a position weight matrix from Equation 3.3.

3.11 Supplementary Information: Additional data and analysis for *yebG*, *purT*, *xylE*, and *dgoR*

yebG

The *yebG* promoter is among a variety of genes known to increase expression when cells are under DNA damage stress (Wade, 2005), and shared the intergenic

region with the *purT* promoter. We see a similar case discussed in the context of the Reg-Seq project and discussed in B.8. In Chapter 3 we considered the *yebG* promoter in cells grown in standard M9 minimal media with 0.5% glucose (Fig. 3.8). While the information footprints and expression shifts appeared to align with annotated binding sites for LexA (which acts as a repressor and so will have a positive expression shift), and the RNAP binding site (negative shift), we did not show evidence for the identity of each binding protein in the main text. Here we present results from our inference of energy matrices using our Sort-Seq data, which confirm the identity of the binding proteins. We also explore the regulation of *yebG* by perturbing the regulatory state through induction of the SOS response (Lomba et al., 1997; Wade, 2005). We begin by considering the Sort-Seq data from cells grown in M9 minimal media with 0.5% glucose. In Fig. 3.10 we show the inferred energy matrices associated with the annotated site for LexA. This was in excellent agreement with the known sequence specificity of LexA (see Fig. 3.4 for a direct comparison with the genomic sequence logos). We note, however, that the RNAP binding site was shifted by 9 bp from the annotated binding site (M. R. Lomba et al., 2006), with an overlap between the -10 RNAP site and 4 bp of the LexA binding site. We were also interested in confirming that the *yebG* promoter responds DNA stress and is induced as part of the SOS response. By repeating Sort-Seq in cells grown in non-lethal concentrations of mitomycin C (1 $\mu\text{g/ml}$) (Lomba et al., 1997) we observed a dramatic increase in expression relative to growth without mitomycin C. Fluorescence histograms showing expression from our plasmid reporter in non-mutagenized promoter constructs are shown in Fig. 3.10(B). From the expression shift plots and information footprint in Fig. 3.10 we find that this is due to loss of repression at the LexA binding site. This is consistent with the expectation that LexA undergoes proteolysis as part of the SOS response (Wade, 2005).

The *purT* promoter

When cells were grown in the presence of adenine, we identified a putative repressor site between the -10 and -35 regions of the RNAP binding site of the *purT* promoter. In our initial attempt to identify the associated transcription factor we performed a DNA affinity purification using conditions that matched the growth conditions where repression was observed. However, as shown in Fig. 3.10 (C), the most significantly enriched protein (GlpR) only showed an enrichment of about 2.9, which was near the shaded region associated with most other non-specific proteins detected. Only upon repeating our purification in the presence of hypoxanthine (10 $\mu\text{g/ml}$) did we

find enrichment of PurR (approximately 350 fold enrichment).

xylE

In the main text it was noted that we could not perform Sort-Seq on the *xylE* promoter unless cells were grown in xylose. In Supplemental Fig. 3.10 (E), we show the associated fluorescence histograms from libraries grown in either glucose or xylose. Interestingly, each mutated window was essentially identical to autofluorescence when cells were grown in glucose. In contrast, growth in xylose showed differential expression for each of the mutated regions. While the promoter was expected to be sensitive to the presence of xylose (Schmidt et al., 2016), this was still a non-obvious result without prior knowledge of whether repressors or activators were involved.

In our analysis we also noted that the identified set of activator binding sites conformed well with the two other promoters regulated by XylR and CRP, namely *xylFG* and *xylAB*. Here we scanned our inferred energy weight matrix across the intergenic regions of *xylFG* and *xylAB*, in order gain further confidence that the identified feature matched the known binding specificity of these transcription factors. These are shown in Fig. 3.10(F). At each position in these plots, we use the energy matrix to calculate the binding energy of the putative transcription factors. For each we identify a strong peak that does indeed align well with the annotated binding sites of XylR and CRP. While our predicted binding energies are not in absolute k_bT units, they are much more negative than the promoter background and predict a similar binding energy (in arbitrary units) to the binding site region of the *xylE* promoter.

dgoR

The last promoter we considered was associated with the expression of the *dgoRKADT* operon. Due to the complexity observed, we were unable to show all data in the main text that supported our identification of the regulatory architecture. In particular, here we show the sensitivity to the different carbon sources considered and additional analysis of the identified regulatory binding sites for DgoR, RNAP, and CRP.

We confirmed that galactose and D-galactonate induce the operon (Cooper, 1978; Schmidt et al., 2016) using fluorescence measurements. In Fig. 3.11 (A) we show information footprints and expression shifts under growth in M9, galactose, and D-galactonate. Additionally, 3 total RNAP sites are present in this promoter, each overlapping. Their sequence logos can be found in 3.13 (B).

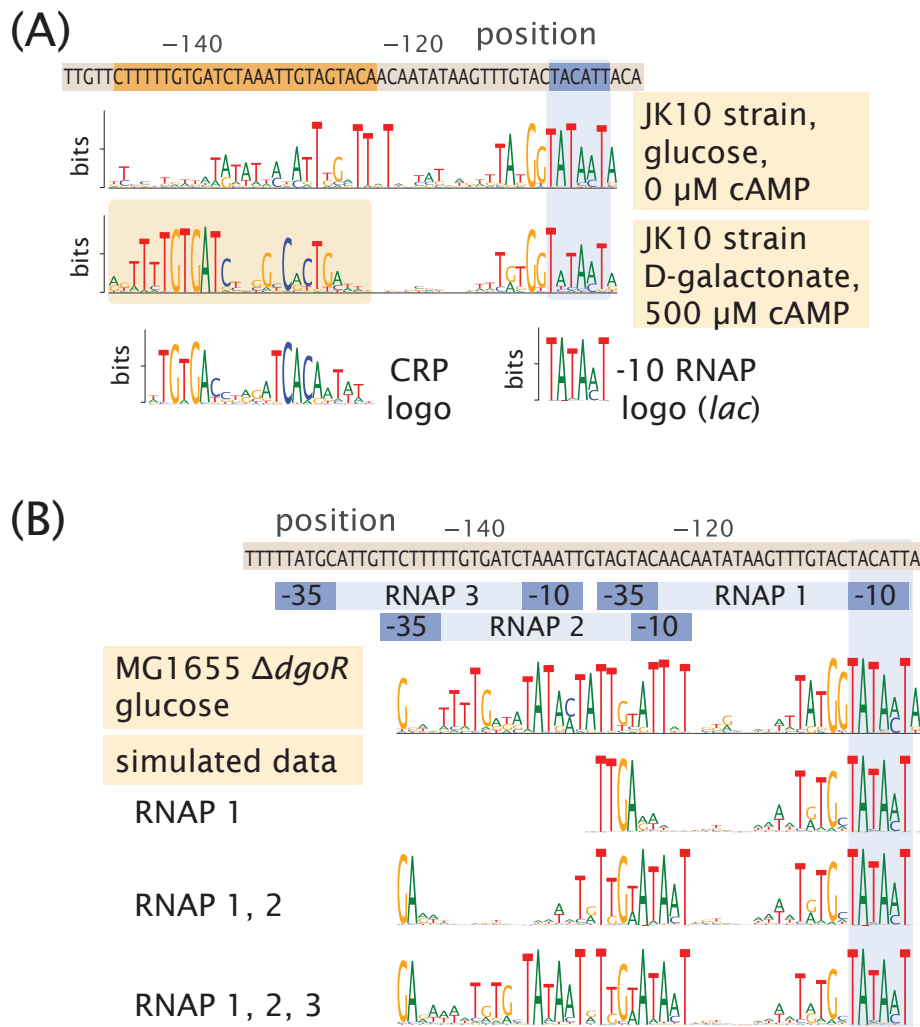


Figure 3.13: Extended analysis of the *dgoR* promoter. (A) Sequence logos were generated for the most upstream 60 bp region containing the putative RNAP and CRP binding sites. Data is from Sort-Seq in strain JK10 (derivative of TK310 (Kinney and Callan, 2010)) and binding of CRP was induced through addition of 500 μ M cAMP. Cells were grown in EZrich MOPS media (Teknova, CA, USA) with D-galactonate as the carbon source. In comparison to the sequence logos shown in Fig. 3.11 (D), the right side of the CRP binding site has become more apparent. (B) Sequence logos are shown for simulated data for the upstream region of the *dgoR* promoter assuming one, two, or three RNAP binding sites. The top sequence logo shows the experimental result for Sort-Seq performed in a Δ *dgoR* genetic background, with cells grown in glucose.

We also show additional evidence to support the claim of a putative binding site for CRP. CRP binds to DNA by coactivation with cAMP. cAMP levels are low in glucose containing media, and at a reasonable level in other growth media. cAMP levels are usually regulated to be insensitive to external sources of cAMP, but, to further enhance CRP binding we used the strain JK10 (based on TK310 Kinney et al., 2010; MG1655 $\Delta cyaA\Delta cpdA$), where we can directly control cAMP to by simple adding it to the growth media. Here we grew cells in EZrich MOPS media (Teknova, CA, USA) with D-galactonate as the carbon source and supplemented with 500 μ M cAMP.

The match from the Sort-Seq data to a genomic CRP binding site, as seen in Fig 3.13(A) is enhanced under these high cAMP levels, further supporting that this is a CRP binding site. This binding site is a class II activator. In the case of lower cAMP levels, it is likely the binding signature is obscured by the presence of the RNAP -35 site.

BIBLIOGRAPHY

- Alekshun, M N and Levy (Oct. 1997). “Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon.” In: *Antimicrobial Agents and Chemotherapy* 41.10, pp. 2067–2075.
- Aono, Rikizo, Norihiko Tsukagoshi, and Mami Yamamoto (1998). “Involvement of Outer Membrane Protein TolC, a Possible Member of the mar-sox Regulon, in Maintenance and Improvement of Organic Solvent Tolerance of Escherichia coli K-12”. en. In: *Journal of Bacteriology* 180.4, pp. 938–944. doi: 10.1128/JB.180.4.938-944.1998.
- Arnold, Cosmas D. et al. (Mar. 2013). “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq”. en. In: *Science* 339.6123, pp. 1074–1077. doi: 10.1126/science.1232542.
- Baba, Tomoya et al. (Jan. 2006). “Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection”. en. In: *Molecular Systems Biology* 2.1. doi: 10.1038/msb4100050.
- Bech, F. W. et al. (Apr. 1985). “Sequence of the relB transcription unit from Escherichia coli and identification of the relB gene.” en. In: *The EMBO Journal* 4.4, pp. 1059–1066. doi: 10.1002/j.1460-2075.1985.tb03739.x.
- Berg, O. G. and P. H. von Hippel (Feb. 1987). “Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters”. eng. In: *Journal of Molecular Biology* 193.4, pp. 723–750. doi: 10.1016/0022-2836(87)90354-8.
- Bintu, Lacramioara et al. (Apr. 2005). “Transcriptional regulation by the numbers: models”. en. In: *Current Opinion in Genetics & Development*. Chromosomes and expression mechanisms 15.2, pp. 116–124. doi: 10.1016/j.gde.2005.02.007.
- Bonocora, Richard P. and Joseph T. Wade (2015). “ChIP-Seq for Genome-Scale Analysis of Bacterial DNA-Binding Proteins”. en. In: *Bacterial Transcriptional Control: Methods and Protocols*. Ed. by Irina Artsimovitch and Thomas J. Santangelo. Methods in Molecular Biology. New York, NY: Springer, pp. 327–340. doi: 10.1007/978-1-4939-2392-2_20.
- Browning, Douglas F and Busby (2016). “Local and global regulation of transcription initiation in bacteria”. In: *Nature Reviews Microbiology*, pp. 638–650. doi: 10.1038/nrmicro.2016.103.
- Busby and Richard H Ebright (Oct. 1999). “Transcription activation by catabolite activator protein (CAP)”. en. In: *Journal of Molecular Biology* 293.2, pp. 199–213. doi: 10.1006/jmbi.1999.3161.
- Cataudella, Ilaria, Kim Sneppen, et al. (Aug. 2013). “Conditional Cooperativity of Toxin - Antitoxin Regulation Can Mediate Bistability between Growth and Dormancy”. In: *PLoS Computational Biology* 9.8. doi: 10.1371/journal.pcbi.1003174.

- Cataudella, Ilaria, Ala Trusina, et al. (Aug. 2012). “Conditional cooperativity in toxin–antitoxin regulation prevents random toxin activation and promotes fast translational recovery”. en. In: *Nucleic Acids Research* 40.14, pp. 6424–6434. DOI: 10.1093/nar/gks297.
- Choi, Kang Yell and Howard Zalkin (1992). “Structural Characterization and Corepressor Binding of the Escherichia coli Purine Repressor”. en. In: *J. BACTERIOL.*, p. 8.
- Cho et al. (Aug. 2011). “The PurR regulon in Escherichia coli K-12 MG1655”. en. In: *Nucleic Acids Research* 39.15, pp. 6456–6464. DOI: 10.1093/nar/gkr307.
- Cipriano, Michael J. et al. (Apr. 2013). “RegTransBase – a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes”. In: *BMC Genomics* 14.1, p. 213. DOI: 10.1186/1471-2164-14-213.
- Cooper, Ronald A. (1978). “The utilisation of d-galactonate and d-2-oxo-3-deoxygalactonate by Escherichia coli K-12: Biochemical and genetical studies”. en. In: *Archives of Microbiology* 118.2, pp. 199–206. DOI: 10.1007/BF00415730.
- Cox, Jürgen and Mann (Dec. 2008). “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification”. en. In: *Nature Biotechnology* 26.12, pp. 1367–1372. DOI: 10.1038/nbt.1511.
- Cox, Jürgen et al. (May 2009). “A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics”. en. In: *Nature Protocols* 4.5, pp. 698–705. DOI: 10.1038/nprot.2009.36.
- Datsenko, Kirill and Barry L. Wanner (June 2000). “One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products”. en. In: *Proceedings of the National Academy of Sciences* 97.12, pp. 6640–6645. DOI: 10.1073/pnas.120163297.
- Fulco, Charles P. et al. (Nov. 2016). “Systematic mapping of functional enhancer–promoter connections with CRISPR interference”. en. In: *Science* 354.6313, pp. 769–773. DOI: 10.1126/science.aag2445.
- Gama-Castro, Socorro et al. (Jan. 2016). “RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond”. en. In: *Nucleic Acids Research* 44.D1, pp. D133–D143. DOI: 10.1093/nar/gkv1156.
- Garcia and Phillips (July 2011). “Quantitative dissection of the simple repression input-output function”. en. In: *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–12178. DOI: 10.1073/pnas.1015616108.
- Gibson, Daniel G et al. (May 2009). “Enzymatic assembly of DNA molecules up to several hundred kilobases”. en. In: *Nature Methods* 6.5, pp. 343–345. DOI: 10.1038/nmeth.1318.

- Gotfredsen, Marie and Kenn Gerdes (1998). “The Escherichia coli relBE genes belong to a new toxin–antitoxin gene family”. en. In: *Molecular Microbiology* 29.4, pp. 1065–1076. DOI: 10.1046/j.1365-2958.1998.00993.x.
- Grainger, David C. et al. (Dec. 2005). “Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome”. en. In: *Proceedings of the National Academy of Sciences* 102.49, pp. 17693–17698. DOI: 10.1073/pnas.0506687102.
- Griffin, Meghan, Jared H. Davis, and Scott A. Strobel (Dec. 2013). “Bacterial Toxin RelE: A Highly Efficient Ribonuclease with Exquisite Substrate Specificity Using Atypical Catalytic Residues”. In: *Biochemistry* 52.48, pp. 8633–8642. DOI: 10.1021/bi401325c.
- Ireland, William T. and Kinney (May 2016). “MPAthic: Quantitative Modeling of Sequence-Function Relationships for massively parallel assays”. en. In: DOI: 10.1101/054676.
- Kalli, Anastasia and Sonja Hess (2012). “Effect of mass spectrometric parameters on peptide and protein identification rates for shotgun proteomic experiments on an LTQ-orbitrap mass analyzer”. en. In: *PROTEOMICS* 12.1, pp. 21–31. DOI: 10.1002/pmic.201100464.
- Keseler, Ingrid M. et al. (Jan. 2013). “EcoCyc: fusing model organism databases with systems biology”. en. In: *Nucleic Acids Research* 41.D1, pp. D605–D612. DOI: 10.1093/nar/gks1027.
- Kheradpour, P. et al. (May 2013). “Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay”. en. In: *Genome Research* 23.5, pp. 800–811. DOI: 10.1101/gr.144899.112.
- Kinney and Atwal (Jan. 2014). “Parametric Inference in the Large Data Limit Using Maximally Informative Models”. In: *Neural Computation* 26.4, pp. 637–653. DOI: 10.1162/NECO_a_00568.
- Kinney and Callan (May 2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. en. In: *Proceedings of the National Academy of Sciences* 107.20, pp. 9158–9163. DOI: 10.1073/pnas.1004290107.
- Kinney, Anand Murugan, et al. (2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.20, pp. 9158–9163. DOI: 10.1073/pnas.1004290107.
- Kılıç, Sefa et al. (Jan. 2014). “CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria”. en. In: *Nucleic Acids Research* 42.D1, pp. D156–D160. DOI: 10.1093/nar/gkt1123.

- Kosuri, Sriram et al. (Aug. 2013). “Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*”. en. In: *Proceedings of the National Academy of Sciences* 110.34, pp. 14024–14029. doi: 10.1073/pnas.1301301110.
- Kuhlman and Edward C. Cox (Apr. 2010). “Site-specific chromosomal integration of large synthetic constructs”. en. In: *Nucleic Acids Research* 38.6, e92–e92. doi: 10.1093/nar/gkp1193.
- Kuhlman, Z. Zhang, et al. (Apr. 2007). “Combinatorial transcriptional control of the lactose operon of *Escherichia coli*”. en. In: *Proceedings of the National Academy of Sciences* 104.14, pp. 6043–6048. doi: 10.1073/pnas.0606717104.
- Laikova, O. N., A. A. Mironov, and M. S. Gelfand (Dec. 2001). “Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria”. eng. In: *FEMS microbiology letters* 205.2, pp. 315–322. doi: 10.1111/j.1574-6968.2001.tb10966.x.
- Lässig, Michael (Sept. 2007). “From biophysics to evolutionary genetics: statistical aspects of gene regulation”. en. In: *BMC Bioinformatics* 8.6, S7. doi: 10.1186/1471-2105-8-S6-S7.
- Latif, Haythem et al. (May 2018). “ChIP-exo interrogation of Crp, DNA, and RNAP holoenzyme interactions”. In: *PLoS ONE* 13.5. doi: 10.1371/journal.pone.0197272.
- Lee, Jae Ok, Cho, and Ok Bin Kim (Oct. 2014). “Overproduction of AcrR increases organic solvent tolerance mediated by modulation of SoxS regulon in *Escherichia coli*”. en. In: *Applied Microbiology and Biotechnology* 98.20, pp. 8763–8773. doi: 10.1007/s00253-014-6024-9.
- Li, Guang-Yao et al. (June 2008). “Structural Mechanism of Transcriptional Autorepression of the *Escherichia coli* RelB/RelE Antitoxin/Toxin Module”. en. In: *Journal of Molecular Biology* 380.1, pp. 107–119. doi: 10.1016/j.jmb.2008.04.039.
- Li et al. (Apr. 2014). “Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources”. en. In: *Cell* 157.3, pp. 624–635. doi: 10.1016/j.cell.2014.02.033.
- Lomba, Mariza R et al. (Jan. 2006). “Identification of yebG as a DNA damage-inducible *Escherichia coli* gene”. en. In: *FEMS Microbiology Letters* 156.1, pp. 119–122. doi: 10.1111/j.1574-6968.1997.tb12715.x.
- Lomba et al. (Nov. 1997). “Identification of yebG as a DNA damage-inducible *Escherichia coli* gene”. en. In: *FEMS Microbiology Letters* 156.1, pp. 119–122. doi: 10.1111/j.1574-6968.1997.tb12715.x.
- Loomis, William F. and Boris Magasanik (Feb. 1967). “The catabolite repression gene of the Lac operon in *Escherichia coli*”. en. In: *Journal of Molecular Biology* 23.3, pp. 487–494. doi: 10.1016/S0022-2836(67)80120-7.

- Lutz, R (Mar. 1997). “Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements”. en. In: *Nucleic Acids Research* 25.6, pp. 1203–1210. DOI: 10.1093/nar/25.6.1203.
- Magoc and Salzberg (Nov. 2011). “FLASH: fast length adjustment of short reads to improve genome assemblies”. en. In: *Bioinformatics* 27.21, pp. 2957–2963. DOI: 10.1093/bioinformatics/btr507.
- Maisonneuve, Etienne and Kenn Gerdes (Apr. 2014). “Molecular Mechanisms Underlying Bacterial Persisters”. en. In: *Cell* 157.3, pp. 539–548. DOI: 10.1016/j.cell.2014.02.050.
- Marbach, Daniel et al. (Aug. 2012). “Wisdom of crowds for robust gene network inference”. en. In: *Nature Methods* 9.8, pp. 796–804. DOI: 10.1038/nmeth.2016.
- Maricque, Brett B., Joseph D. Dougherty, and Barak A. Cohen (Oct. 2016). “A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of *cis* -regulatory activity in neural cells”. en. In: *Nucleic Acids Research*, gkw942. DOI: 10.1093/nar/gkw942.
- Martin, R G and J L Rosner (1997). “Fis, an accessorial factor for transcriptional activation of the *mar* (multiple antibiotic resistance) promoter of *Escherichia coli* in the presence of the activator MarA, SoxS, or Rob.” en. In: *Journal of bacteriology* 179.23, pp. 7410–7419. DOI: 10.1128/JB.179.23.7410-7419.1997.
- Melnikov, Alexandre et al. (Mar. 2012). “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay”. en. In: *Nature Biotechnology* 30.3, pp. 271–277. DOI: 10.1038/nbt.2137.
- Minchin, Stephen D. and Busby (Jan. 2009). “Analysis of mechanisms of activation and repression at bacterial promoters”. en. In: *Methods. Methods Related to Bacterial Transcriptional Control* 47.1, pp. 6–12. DOI: 10.1016/j.ymeth.2008.10.012.
- Mirzaei, Hamid et al. (Feb. 2013). “Systematic measurement of transcription factor-DNA interactions by targeted mass spectrometry identifies candidate gene regulatory proteins”. en. In: *Proceedings of the National Academy of Sciences* 110.9, pp. 3645–3650. DOI: 10.1073/pnas.1216918110.
- Mittler, G., F. Butter, and M. Mann (Dec. 2008). “A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements”. en. In: *Genome Research* 19.2, pp. 284–293. DOI: 10.1101/gr.081711.108.
- (2009). “A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements”. In: *Genome Res* 19.2, pp. 284–93. DOI: 10.1101/gr.081711.108.

- Munch (Jan. 2003). “PRODORIC: prokaryotic database of gene regulation”. en. In: *Nucleic Acids Research* 31.1, pp. 266–269. DOI: 10.1093/nar/gkg037.
- Nishida, Keishin, Martin C. Frith, and Kenta Nakai (Feb. 2009). “Pseudocounts for transcription factor binding sites”. en. In: *Nucleic Acids Research* 37.3, pp. 939–944. DOI: 10.1093/nar/gkn1019.
- Oehler, S. (Jan. 2006). “Induction of the lac promoter in the absence of DNA loops and the stoichiometry of induction”. en. In: *Nucleic Acids Research* 34.2, pp. 606–612. DOI: 10.1093/nar/gkj453.
- Oehler et al. (Apr. 1990). “The three operators of the lac operon cooperate in repression.” en. In: *The EMBO Journal* 9.4, pp. 973–979. DOI: 10.1002/j.1460-2075.1990.tb08199.x.
- Okuda, Shujiro et al. (Jan. 2017). “jPOSTrepo: an international standard data repository for proteomes”. en. In: *Nucleic Acids Research* 45.D1, pp. D1107–D1111. DOI: 10.1093/nar/gkw1080.
- Ong, Shao-En et al. (May 2002). “Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics”. en. In: *Molecular & Cellular Proteomics* 1.5, pp. 376–386. DOI: 10.1074/mcp.M200025-MCP200.
- Overgaard, Martin, Jonas Borch, and Kenn Gerdes (Nov. 2009). “RelB and RelE of Escherichia coli Form a Tight Complex That Represses Transcription via the Ribbon–Helix–Helix Motif in RelB”. en. In: *Journal of Molecular Biology* 394.2, pp. 183–196. DOI: 10.1016/j.jmb.2009.09.006.
- Overgaard, Martin, Jonas Borch, Mikkel G. Jørgensen, et al. (2008). “Messenger RNA interferase RelE controls relBE transcription by conditional cooperativity”. en. In: *Molecular Microbiology* 69.4, pp. 841–857. DOI: 10.1111/j.1365-2958.2008.06313.x.
- Patwardhan, Rupali P et al. (Feb. 2012). “Massively parallel functional dissection of mammalian enhancers in vivo”. In: *Nature biotechnology* 30.3, pp. 265–270. DOI: 10.1038/nbt.2136.
- Rappsilber, Juri, Mann, and Yasushi Ishihama (Aug. 2007). “Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips”. en. In: *Nature Protocols* 2.8, pp. 1896–1906. DOI: 10.1038/nprot.2007.261.
- Razo-Mejia, M. et al. (Apr. 2014). “Comparison of the theoretical and real-world evolutionary potential of a genetic circuit”. In: *Physical biology* 11.2, p. 026005. DOI: 10.1088/1478-3975/11/2/026005.
- Rolfes (Oct. 2006). “Regulation of purine nucleotide biosynthesis: in yeast and beyond”. en. In: *Biochemical Society Transactions* 34.5, pp. 786–790. DOI: 10.1042/BST0340786.

- Ruiz, C. and Levy (May 2010). “Many Chromosomal Genes Modulate MarA-Mediated Multidrug Resistance in *Escherichia coli*”. en. In: *Antimicrobial Agents and Chemotherapy* 54.5, pp. 2125–2134. DOI: 10.1128/AAC.01420-09.
- Sawitzke, James et al. (Jan. 2007). “Recombineering: In Vivo Genetic Engineering in *E. coli*, *S. enterica*, and Beyond”. en. In: *Methods in Enzymology*. Vol. 421. Advanced Bacterial Genetics: Use of Transposons and Phage for Genomic Engineering. Academic Press, pp. 171–199. DOI: 10.1016/S0076-6879(06)21015-2.
- Schmidt, Alexander et al. (Jan. 2016). “The quantitative and condition-dependent *Escherichia coli* proteome”. en. In: *Nature Biotechnology* 34.1, pp. 104–110. DOI: 10.1038/nbt.3418.
- Schneider, T D and R M Stephens (Oct. 1990). “Sequence logos: a new way to display consensus sequences.” In: *Nucleic Acids Research* 18.20, pp. 6097–6100.
- Seoane, A S and Levy (1995). “Characterization of MarR, the repressor of the multiple antibiotic resistance (*mar*) operon in *Escherichia coli*.” en. In: *Journal of bacteriology* 177.12, pp. 3414–3419. DOI: 10.1128/JB.177.12.3414-3419.1995.
- Sharon, Eilon et al. (June 2012). “Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters”. en. In: *Nature Biotechnology* 30.6, pp. 521–530. DOI: 10.1038/nbt.2205.
- Shimada, K. Yamamoto, and A. Ishihama (Feb. 2011). “Novel Members of the *Cra* Regulon Involved in Carbon Metabolism in *Escherichia coli*”. en. In: *Journal of Bacteriology* 193.3, pp. 649–659. DOI: 10.1128/JB.01214-10.
- Singh et al. (Feb. 2014). “Widespread suppression of intragenic transcription initiation by H-NS”. en. In: *Genes & Development* 28.3, pp. 214–219. DOI: 10.1101/gad.234336.113.
- Song, S and C Park (1997). “Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator.” en. In: *Journal of bacteriology* 179.22, pp. 7025–7032. DOI: 10.1128/JB.179.22.7025-7032.1997.
- Stormo, G. D. (Jan. 2000). “DNA binding sites: representation and discovery”. en. In: *Bioinformatics* 16.1, pp. 16–23. DOI: 10.1093/bioinformatics/16.1.16.
- Thomason, Lynn C., Nina Costantino, and Donald L. Court (2007). “*E. coli* Genome Manipulation by P1 Transduction”. en. In: *Current Protocols in Molecular Biology* 79.1, pp. 1.17.1–1.17.8. DOI: 10.1002/0471142727.mb0117s79.
- Vvedenskaya, Irina O., Seth Goldman, and Bryce E. Nickels (2015). “Preparation of cDNA libraries for high-throughput RNA sequencing analysis of RNA 5 ends”. In: *Methods in molecular biology (Clifton, N.J.)* 1276, pp. 211–228. DOI: 10.1007/978-1-4939-2392-2_12.

- Wade (Nov. 2005). “Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia coli* genome and identifies unconventional target sites”. en. In: *Genes & Development* 19.21, pp. 2619–2630. DOI: 10.1101/gad.1355605.
- Weatherspoon-Griffin, Natasha et al. (Nov. 2014). “The CpxR/CpxA Two-component Regulatory System Up-regulates the Multidrug Resistance Cascade to Facilitate *Escherichia coli* Resistance to a Model Antimicrobial Peptide”. en. In: *Journal of Biological Chemistry* 289.47, pp. 32571–32582. DOI: 10.1074/jbc.M114.565762.
- Wiśniewski, Jacek R et al. (May 2009). “Universal sample preparation method for proteome analysis”. en. In: *Nature Methods* 6.5, pp. 359–362. DOI: 10.1038/nmeth.1322.
- Yamaguchi, Yoshihiro and Masayori Inouye (Nov. 2011). “Regulation of growth and death in *Escherichia coli* by toxin–antitoxin systems”. en. In: *Nature Reviews Microbiology* 9.11, pp. 779–790. DOI: 10.1038/nrmicro2651.
- Zhang, Huibin et al. (Apr. 2016). “Comprehensive mutagenesis of the fimS promoter regulatory switch reveals novel regulation of type 1 pili in uropathogenic *Escherichia coli*”. en. In: *Proceedings of the National Academy of Sciences* 113.15, pp. 4182–4187. DOI: 10.1073/pnas.1522958113.
- Zheng, Dongling et al. (Oct. 2004). “Identification of the CRP regulon using in vitro and in vivo transcriptional profiling”. en. In: *Nucleic Acids Research* 32.19, pp. 5874–5893. DOI: 10.1093/nar/gkh908.

Chapter 4

DECIPHERING THE REGULATORY GENOME OF *ESCHERICHIA COLI*, ONE HUNDRED PROMOTERS AT A TIME

A version of this chapter originally appeared as W. T. Ireland, S. M. Beeler, E. Flores-Bautista, N. M. Belliveau, M. J. Sweredoski, J. B. Kinney, and R. Phillips (2020). Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time. bioRxiv, <http://doi.org/10.1101/2020.01.18.910323>.

Author contribution note: for this chapter, I (WI) assisted with experimental design, sample processing, strain construction, RNA-seq measurements, data analysis, mass-spectrometry measurements, and manuscript writing.

4.1 Introduction

DNA sequencing is as important to biology as the telescope is to astronomy. As discussed in previous chapters, we are now living in the age of genomics, where DNA sequencing has become cheap and routine. However, despite these incredible advances, how all of this genomic information is regulated and deployed remains largely enigmatic. Organisms must respond to their environments through regulation of genes. Genomic methods often provide a “parts” list but often leave us uncertain about how those parts are used creatively and constructively in space and time. Yet, we know that promoters apply all-important dynamic logical operations that control when and where genetic information is accessed. In this chapter, we demonstrate how we can infer the logical and regulatory interactions that control bacterial decision making across large numbers of promoters by tapping into the power of DNA sequencing as a biophysical tool. The method introduced here provides a framework for solving the problem of deciphering the regulatory genome by connecting perturbation and response, mapping information flow from individual nucleotides in a promoter sequence to downstream gene expression, determining how much information each promoter base pair carries about the level of gene expression.

The advent of RNA-Seq (Mortazavi et al., 2008) launched a new era in which se-

quencing could be used as an experimental read-out of the biophysically interesting counts of mRNA, rather than simply as a tool for collecting ever more complete organismal genomes. The slew of ‘X’-Seq technologies that are available continues to expand at a dizzying pace, each serving their own creative and insightful role: RNA-Seq, ChIP-Seq, Tn-Seq, SELEX, 5C, etc. (Stuart and Satija, 2019). In contrast to whole genome screening sequencing approaches, such as Tn-Seq (Goodall et al., 2018) and ChIP-Seq (Gao et al., 2018) which give a coarse-grained view of gene essentiality and regulation respectively, another class of experiments known as massively-parallel reporter assays (MPRA) has been used to study gene expression in a variety of contexts (Patwardhan, Lee, et al., 2009; Kinney, Murugan, et al., 2010; Sharon et al., 2012; Patwardhan, Hiatt, et al., 2012; Melnikov et al., 2012; Kwasniewski et al., 2012; Fulco et al., 2019; Kinney and McCandlish, 2019). One elegant study relevant to the bacterial case of interest here by Kosuri et al., 2013 screened more than 10^4 combinations of promoter and ribosome binding sites (RBS). Even more recently, they have utilized MPRA in sophisticated ways to search for regulated genes across the genome (G. Urtecho et al., 2019; Guillaume Urtecho et al., 2020), in a way we see as being complementary to our own. While their approach yields a coarse-grained view of where regulation may be occurring, our approach yields a base-pair-by-base-pair view of how exactly that regulation is being enacted.

One of the most exciting X-Seq tools based on MPRA with broad biophysical reach is the Sort-Seq approach developed by Kinney, Murugan, et al., 2010, with proof of concept work on virgin genes discussed in the previous chapter. Sort-Seq uses fluorescence activated cell sorting (FACS) based on changes in the fluorescence due to mutated promoters to identify the specific locations of transcription factor binding in the genome. Importantly, it also provides a readout of how promoter sequences control the level of gene expression with single base-pair resolution. The results of such a massively-parallel reporter assay make it possible to build a biophysical model of gene regulation to uncover how previously uncharacterized promoters are regulated. In particular, high-resolution studies like those described here yield quantitative predictions about promoter organization and protein-DNA interactions as described by energy matrices (Kinney, Murugan, et al., 2010). This allows us to employ the tools of statistical physics to describe the input-output properties of each of these promoters which can be explored much further with in-depth experimental dissection like those done by Razo-Mejia et al., 2018 and Chure et al., 2019 and summarized in Phillips et al., 2019. In this sense, the Sort-Seq approach can provide a quantitative

framework to not only discover and quantitatively dissect regulatory interactions at the promoter level, but also provides an interpretable scheme to design genetic circuits with a desired expression output as discussed in Chapter 2 (Barnes et al., 2019).

Earlier work discussed in Chapter 3 (Belliveau et al., 2018) illustrated how Sort-Seq, used in conjunction with mass spectrometry can be used to identify which transcription factors bind to a given binding site, thus enabling the mechanistic dissection of promoters which previously had no regulatory annotation. However, a crucial drawback of the Sort-Seq approach is that while it is high-throughput at the level of a single gene and the number of promoter variants it accesses, it was unable to readily tackle multiple genes at once, still leaving much of the unannotated genome untouched. Given that even in one of biology's best understood organisms, the bacterium *Escherichia coli*, for more than 65% of its genes, we remain completely ignorant of how those genes are regulated (Santos-Zavaleta et al., 2019; Belliveau et al., 2018). If we hope to some day have a complete base pair resolution mapping of how genetic sequences relate to biological function, we must first be able to do so for the promoters of this "simple" organism.

What has been missing in uncovering the regulatory genome in organisms of all kinds is a large scale method for inferring genomic logic and regulation. Here we replace the low-throughput fluorescence-based Sort-Seq approach with a scalable RNA-Seq based approach that makes it possible to attack multiple promoters at once, setting the stage for the possibility of, to first approximation, uncovering the entirety of the regulatory genome. Accordingly, we refer to the entirety of our approach (MPRA, information footprints and energy matrices, mass spectrometry for transcription factor identification) as Reg-Seq, which we employ here on over one hundred promoters. The concept of MPRA methods is to perturb promoter regions by mutating them and then using sequencing to read out both perturbation and the resulting gene expression (Patwardhan, Lee, et al., 2009; Kinney, Murugan, et al., 2010; Sharon et al., 2012; Patwardhan, Hiatt, et al., 2012; Melnikov et al., 2012; Kwasnieski et al., 2012; Fulco et al., 2019; Kinney and McCandlish, 2019). We generate a broad diversity of promoter sequences for each promoter of interest and use mutual information as a metric to measure information flow from that distribution of sequences to gene expression. Thus, Reg-Seq is able to collect causal information about candidate regulatory sequences that is then complemented by

mass spectrometry which allows us to find which transcription factors mediate the action of those newly discovered candidate regulatory sequences. Hence, Reg-Seq solves the causal problem of linking DNA sequence to regulatory logic and information flow.

To demonstrate our ability to scale up Sort-Seq with the sequencing based Reg-Seq protocol, we report here our results for 113 *E. coli* genes, whose regulatory architectures (i.e. gene-by-gene distributions of transcription factor (TF) binding sites and identities of TFs that bind those sites) were determined in parallel. By taking the Sort-Seq approach from a gene-by-gene method to a more whole-genome approach, we can begin to piece together not just how individual promoters are regulated, but also the nature of gene-gene interactions by revealing how certain transcription factors serve to regulate multiple genes at once. This approach has the benefits of a high-throughput assay while sacrificing little of the resolution afforded by the previous gene-by-gene approach, allowing us to uncover a large swath of the *E. coli* regulome, with base-pair resolution, in one set of experiments.

The organization of the remainder of the Chapter is as follows. In the Results section, we provide a global view of the discoveries we made in our exploration of more than 100 promoters in *E. coli* using Reg-Seq. These results are described in summary form in the paper itself, with a full online version of the results (www.rpgroup.caltech.edu/RNAseq_SortSeq/interactive_a) showing how different growth conditions elicit different regulatory responses. This section also follows the overarching view of our results by examining several biological stories that emerge from our data and serve as case studies in what has been revealed in our efforts to uncover the regulatory genome. The Discussion section summarizes the method and the current round of discoveries it has afforded with an eye to future applications to further elucidate the *E. coli* genome and opening up the quantitative dissection of other non-model organisms. In the Methods section and fleshed out further in the Appendices, we describe our methodology and benchmark it against our own earlier Sort-Seq experiments to show that using RNA-Seq as a readout of the expression of mutated promoters is equally reliable as the fluorescence-based approach. Lastly, in the appendices for this chapter we discuss innovations, that while they were not used for the Reg-Seq work discussed here, are on the horizon and will be part of the next generation of Reg-Seq experiments,

such as using neural networks to model DNA-protein interactions, and utilizing our models to perform more accurate computational searches for transcription factor binding sites.

4.2 Results

Selection of genes and methodology

As shown in Figure 4.1, we have considered more than 100 genes from across the *E. coli* genome. Our choices were based on a number of factors (see Section B.1 for more details); namely, we wanted a subset of genes that served as a “gold standard” for which the hard work of generations of molecular biologists have yielded deep insights into their regulation. The set includes *lacZYA*, *znuCB*, *znuA*, *ompR*, *araC*, *marR*, *relBE*, *dgoR*, *dicC*, *ftsK*, *xylA*, *xylF*, *dpiBA*, *rspA*, *dicA*, and *araAB*. By using Reg-Seq on these genes we were able to demonstrate that this method recovers not only what was already known about binding sites and transcription factors for well-characterized promoters, but also whether there are any important differences between the results of the methods presented here and the previous generation of experiments based on fluorescence and cell-sorting as a readout of gene expression. These promoters of known regulatory architecture are complemented by an array of previously uncharacterized genes that we selected in part using data from a recent proteomic study, in which mass spectrometry was used to measure the copy number of different proteins in 22 distinct growth conditions (Schmidt et al., 2016). We selected genes that exhibited a wide variation in their copy number over the different growth conditions considered, reasoning that differential expression across growth conditions implies that those genes are under regulatory control.

As noted in the introduction, the original formulation of Reg-Seq termed Sort-Seq was based on the use of fluorescence activated cell sorting one gene at a time as a way to uncover putative binding sites for previously uncharacterized promoters (Belliveau et al., 2018). As a result, as shown in Figure 4.2 we have formulated a second generation version that permits a high-throughput interrogation of the genome. A comparison between the Sort-Seq and Reg-Seq approaches for the same genes is shown in Figure B.1. In the Reg-Seq approach, for each promoter interrogated, we generate a library of mutated variants and design each variant to express an mRNA with a unique sequence barcode. By counting the frequency of each expressed barcode using RNA-Seq, we can assess the differential expression from our

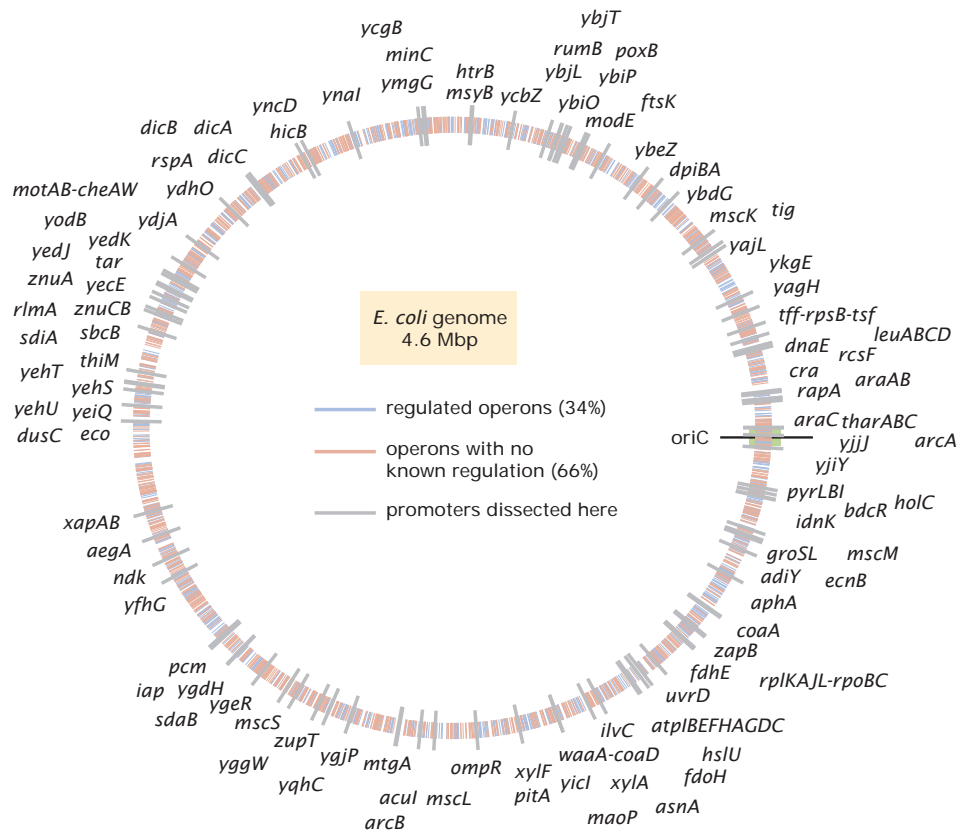


Figure 4.1: The *E. coli* regulatory genome and the genes studied with Reg-Seq. Illustration of the current ignorance with respect to how genes are regulated in *E. coli*, with genes with previously annotated regulation (as reported on RegulonDB (Gama-Castro et al., 2016)) denoted with blue ticks and genes with no previously annotated regulation denoted with red ticks. The 113 genes explored in this study are labeled in gray.

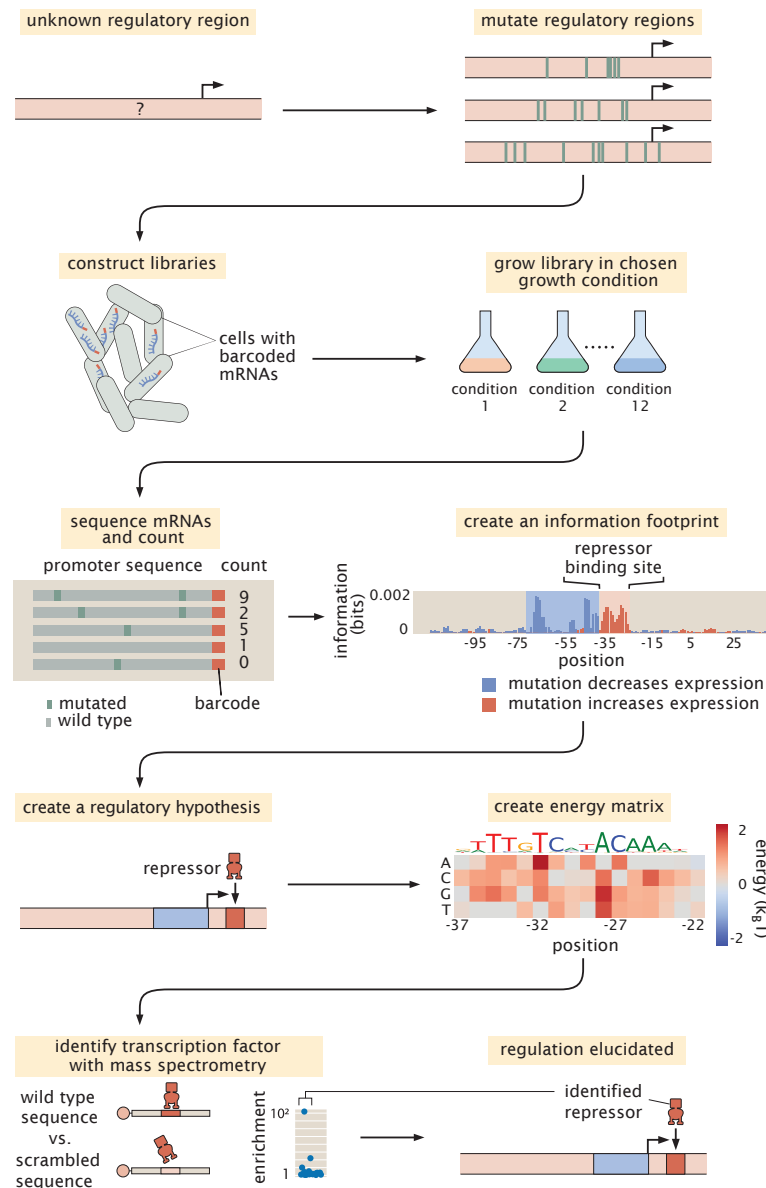


Figure 4.2: The Reg-Seq procedure used to determine how a given promoter is regulated. This process is as follows: After constructing a promoter library driving expression of a randomized barcode (an average of 5 for each promoter), RNA-Seq is conducted to determine frequency of these mRNA barcodes across different growth conditions. By computing the mutual information between DNA sequence and mRNA barcode counts for each base pair in the promoter region, an "information footprint" is constructed yielding a regulatory hypothesis for the putative binding sites. Energy matrices, which describe the effect any given mutation has on DNA binding energy, and sequence logos are inferred for the putative transcription factor binding sites. Next, we identify which transcription factor preferentially binds to the putative binding site via DNA affinity chromatography followed by mass spectrometry. Finally, this procedure culminates in a coarse-grained cartoon-level view of our regulatory hypothesis for how this given promoter is regulated.

promoter of interest based on the base-pair-by-base-pair sequence of its promoter. Using the mutual information between mRNA counts and sequences, we develop an information footprint that reveals the importance of different bases in the promoter region to the overall level of expression. We locate potential transcription factor binding regions by looking for clusters of base pairs that have a significant effect on gene expression. Further details on how potential binding sites are identified are found in the Methods section. Blue regions of the histogram shown in the information footprints of Figure 4.2 correspond to hypothesized activating sequences and red regions of the histogram correspond to hypothesized repressing sequences. With the information footprint in hand, we can then determine energy matrices and sequence logos (described in the next section). Given putative binding sites, we construct oligonucleotides that serve as fishing hooks to fish out the transcription factors that bind to those putative binding sites using DNA-affinity chromatography and mass spectrometry (Mittler, Butter, and M. Mann, 2009). Given all of this information, we can then formulate a schematized view of the newly discovered regulatory architecture of the previously uncharacterized promoter. For the case schematized in Figure 4.2, the experimental pipeline yields a complete picture of a simple repression architecture (i.e. a gene regulated by a single binding site for a repressor).

Visual tools for data presentation

Throughout our investigation of the more than 100 genes explored in this study, we repeatedly relied on several key approaches to help make sense of the immense amount of data generated in these experiments. As these different approaches to viewing the results will appear repeatedly throughout the paper, here we familiarize the reader with five graphical representations referred to respectively as information footprints, energy matrices, sequence logos, mass spectrometry enrichment plots, and regulatory cartoons, which taken all together provide a quantitative description of previously uncharacterized promoters.

Information footprints: From our mutagenized libraries of promoter regions, we can build up a base-pair-by-base-pair graphical understanding of how the promoter sequence relates to level of gene expression in the form of the information footprint shown in the middle of Figure 4.2. In this plot, the bar above each base pair position represents how large of an effect mutations at this location have on the level of

gene expression. Specifically, the quantity plotted is the mutual information I_b at base pair b between mutation of a base pair at that position and the level of expression. In mathematical terms, the mutual information measures how much the joint probability $p(m, \mu)$ differs from the product of the probabilities $p_{mut}(m)p_{expr}(\mu)$ which would be produced if mutation and gene expression level were independent. Formally, the mutual information between having a mutation at position b and level of expression is given by

$$I_b = \sum_{m=0}^1 \sum_{\mu=0}^1 p(m, \mu) \log_2 \left(\frac{p(m, \mu)}{p_{mut}(m)p_{expr}(\mu)} \right). \quad (4.1)$$

Note that both m and μ are binary variables that characterize the mutational state of the base of interest and the level of expression, respectively. Specifically, m can take the values

$$m = \begin{cases} 0, & \text{if } b \text{ is a mutated base} \\ 1, & \text{if } b \text{ is a wild-type base,} \end{cases} \quad (4.2)$$

and μ can take on values

$$\mu = \begin{cases} 0, & \text{for sequencing reads from the DNA library} \\ 1, & \text{for sequencing reads originating from mRNA,} \end{cases} \quad (4.3)$$

where both m and μ are index variables that tell us whether the base has been mutated and if so, how likely that the read at that position will correspond to an mRNA, reflecting gene expression or a promoter, reflecting a member of the library. The higher the ratio of mRNA to DNA reads at a given base position, the higher the expression. $p_{mut}(m)$ in equation 4.1 refers to the probability that a given sequencing read will be from a mutated base. $p_{expr}(\mu)$ is a normalizing factor that gives the ratio of the number of DNA or mRNA sequencing counts to total number of counts.

Furthermore, we color the bars based on whether mutations at this location lowered gene expression on average (in blue, indicating an activating role) or increased gene expression (in red, indicating a repressing role). Within these footprints, we look for regions of approximately 10 to 20 contiguous base pairs which impact gene expression similarly (either increasing or decreasing), as these regions implicate the influence of a transcription factor binding site. In this experiment, we targeted the regulatory regions based on a guess of where a transcription start site (TSS) will be, based on experimentally confirmed sites contained in regulonDB (Santos-Zavaleta

et al., 2019), a 5' RACE experiment (Mendoza-Vargas et al., 2009), or by targeting small intergenic regions. After completing the Reg-Seq experiment, we note that many of the presumed TSS sites are not in the locations assumed, the promoters have multiple active RNA polymerase (RNAP) sites and TSS, or the primary TSS shifts with growth condition. To simplify the data presentation, the "0" base pair in all information footprints is set to the originally assumed base pair for the primary TSS, rather than one of the TSS that was found in the experiment. The locations of the TSS are listed in Table A.1. As can be seen throughout the paper (see Figure 4.4 for several examples of each of the main types of regulatory architectures) and the online resource, we present such information footprints for every promoter we have considered, with one such information footprint for every growth condition.

Energy matrices: Focusing on an individual putative transcription factor binding site as revealed in the information footprint, we are interested in a more fine-grained, quantitative understanding of how the underlying protein-DNA interaction is determined. An energy matrix displays this information using a heat map format, where each column is a position in the putative binding site and each row displays the effect on binding that results from mutating to that given nucleotide (given as a change in the DNA-TF interaction energy upon mutation) (Berg and Hoppel, 1987; Stormo and Fields, 1998; Kinney, Murugan, et al., 2010). These energy matrices are scaled such that the wild type sequence is colored in white, mutations that improve binding are shown in blue, and mutations that weaken binding are shown in red. These energy matrices encode a full quantitative picture for how we expect sequence to relate to binding for a given transcription factor, such that we can provide a prediction for the binding energy of every possible binding site sequence as

$$\text{binding energy} = \sum_{i=1}^N \varepsilon_i, \quad (4.4)$$

where the energy matrix is predicated on an assumption of a linear binding model in which each base within the binding site region contributes a specific value (ε_i for the i^{th} base in the sequence) to the total binding energy. Energy matrices are either given in A.U. (arbitrary units), or if the gene has a simple repression or activation architecture with a single RNA polymerase (RNAP) site, are assigned $k_B T$ energy units following the procedure in Kinney, Murugan, et al., 2010 and validated on the *lac* operon in Barnes et al., 2019.

Sequence logos: From an energy matrix, we can also represent a preferred transcription factor binding site with the use of the letters corresponding to the four possible nucleotides, as is often done with position weight matrices (Schneider and Stephens, 1990). In these sequence logos, the size of the letters corresponds to how strong the preference is for that given nucleotide at that given position, which can be directly computed from the energy matrix. This method of visualizing the information contained within the energy matrix is more easily digested and allows for quick comparison among various binding sites.

Mass spectrometry enrichment plots: As the final piece of our experimental pipeline, we wish to determine the identity of the transcription factor we suspect is binding to our putative binding site that is represented in the energy matrix and sequence logo. While the details of the DNA affinity chromatography and mass spectrometry can be found in the methods, the results of these experiments are displayed in enrichment plots such as is shown in the bottom panel of Figure 4.2. In these plots, the relative abundance of each protein bound to our site of interest is quantified relative to a scrambled control sequence. The putative transcription factor is the one we find to be highly enriched compared to all other DNA binding proteins.

Regulatory cartoons: The ultimate result of all these detailed base-pair-by-base-pair resolution experiments yields a cartoon model of how we think the given promoter is being regulated. A complete set of cartoons for all the architectures considered in our study is presented in Figure 4.9. While the cartoon serves as a convenient visual way to summarize our results, it's important to remember that these cartoons are a shorthand representation of all the data in the four quantitative measures described above and are in fact backed by quantitative predictions of how we expect the system to behave which can be tested experimentally. Throughout this paper we use consistent iconography to illustrate the regulatory architecture of promoters, with activators and their binding sites in green, repressors in red, and RNAP in blue.

Newly discovered *E. coli* regulatory architectures

Figure 4.3 (and Tables 4.1 and 4.2) provides a summary of the discoveries made in the work done here using our next generation Reg-Seq approach. Figure 4.3(A) provides a shorthand notation that conveniently characterizes the different kinds of

regulatory architectures found in bacteria. In previous work (Rydenfelt et al., 2014), we have explored the entirety of what is known about the regulatory genome of *E. coli*, revealing that the most common motif is the (0,0) constitutive architecture, though we hypothesized that this is not a statement about the facts of the *E. coli* genome, but rather a reflection of our collective regulatory ignorance in the sense that we suspect that with further investigation, many of these apparent constitutive architectures will be found to be regulated under the right environmental conditions. The two most common regulatory architectures that emerged from our previous database survey are the (0,1) and (1,0) architectures, the simple repression motif and the simple activation motif, respectively. It is interesting to consider that the (0,1) architecture is in fact the repressor-operator model originally introduced in the early 1960s by Jacob and Monod as the concept of gene regulation emerged (Jacob and Monod, 1961). Now we see retrospectively the far-reaching importance of that architecture across the *E. coli* genome.

For the 113 genes we considered, Figure 4.3(B) summarizes the number of simple repression (0, 1) architectures discovered, the number of simple activation (1, 0) architectures discovered and so on. A comparison of the frequency of the different architectures found in our study to the frequencies of all the known architectures in the RegulonDB database is provided in Figure B.9. Tables 4.1 and 4.2 provide a more detailed view of our results. As seen in Table 4.1, of the 113 genes we considered, 32 of them revealed no signature of any transcription factor binding sites and they are labeled as (0, 0). The simple repression architecture (0, 1) was found 26 times, the simple activation architecture (1, 0) was found 13 times, and more complex architectures featuring multiple binding sites (e.g. (1, 1), (0, 2), (2, 0), etc.) were revealed as well. Further, for 18 of the genes that we label “inactive”, Reg-Seq didn’t even reveal an RNAP binding site. The lack of observable RNAP site could be because the proper growth condition to get high levels of expression was not used, or because the mutation window chosen for the gene does not capture a highly transcribing TSS. The tables also include our set of 16 “gold standard” genes for which previous work has resulted in a knowledge (sometimes only partial) of their regulatory architectures. We find that our method recovers the regulatory elements of these gold standard cases fully in 12 out of 16 cases, and the majority of regulatory elements in 2 of the remainder. Overall the performance of Reg-Seq in these gold-standard cases (for more details see Figure B.2) builds confidence in the approach. Further, the failure modes inform us of the blind spots of Reg-Seq.

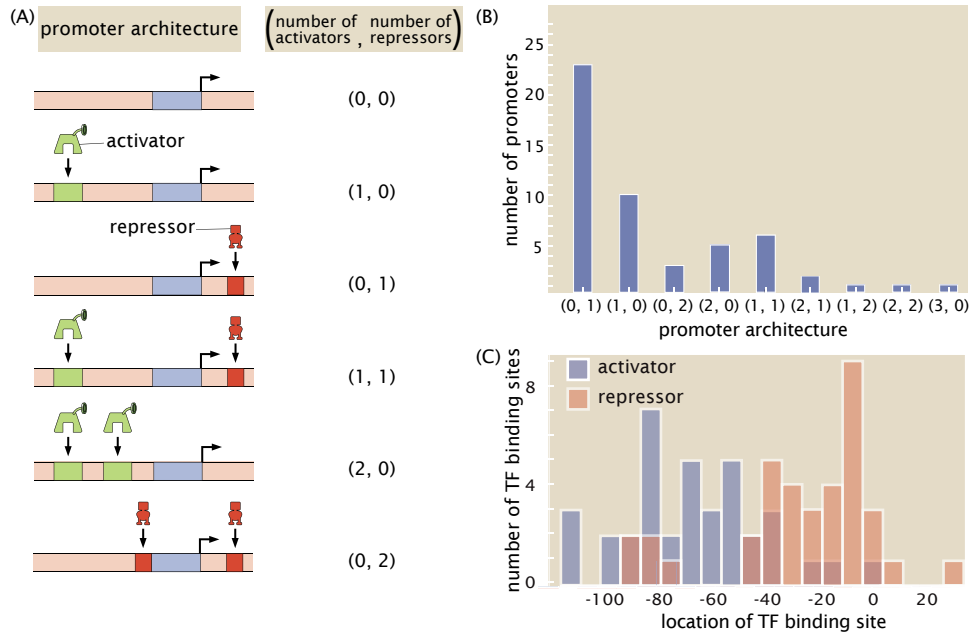


Figure 4.3: A summary of regulatory architectures discovered in this study. (A) The cartoons display a representative example of each type of architecture, along with the corresponding shorthand notation. (B) Counts of the different regulatory architectures discovered in this study. Only those promoters where at least one new binding site was discovered are included in this figure. If one repressor was newly discovered and two activators were previously known, then the architecture is still counted as a (2,1) architecture. (C) Distribution of positions of binding sites discovered in this study for activators and repressors. Only newly discovered binding sites are included in this figure. The position of the TF binding sites are calculated relative to the estimated TSS location, which is based on the location of the associated RNAP site.

For example, we find it challenging to observe weaker binding sites when multiple strong binding sites are also present such as in the *marRAB* operon. Additionally the method will fail when there is no active TSS in the mutation window, as occurred in the case of *dicA*. Further details on the comparison to gold standard genes can be found in section B.1.

We observe that the most common motif to emerge from our work is the simple repression motif. Another relevant regulatory statistic is shown in Figure 4.3(C) where we see the distribution of binding site positions. Our own experience in the use of different quantitative modeling approaches to consider transcriptional regulation reveal that, for now, we remain largely ignorant of how to account for transcription factor binding site position, and datasets like that presented here will

Architecture	Total number of promoters	Number of promoters with at least one newly discovered binding site
All Architectures	113	52
(0,0)	32	0
(0,1)	26	23
(1,0)	13	10
(1,1)	6	6
(0,2)	4	3
(2,0)	6	5
(2,1)	2	2
(1,2)	1	1
(2,2)	1	1
(3,0)	3	1
(0,4)	1	0
inactive	18	0

Table 4.1: All promoters examined in Reg-Seq, categorized according to type of regulatory architecture.

Those promoters which have no recognizable RNAP site are labeled as inactive rather than constitutively expressed (0, 0).

begin to provide data that can help us uncover how this parameter dictates gene expression. Indeed, with binding site positions and energy matrices in hand, we can systematically move these binding sites and explore the implications for the level of gene expression, providing a systematic tool to understand the role of binding-site position.

Figure 4.4 delves more deeply into the various regulatory architectures described in Figure 4.3(B) by showing several example promoters for each of the different architecture types. In each of the cases shown in the figure, prior to the work presented here, these promoters had no regulatory information in relevant databases such as Ecocyc (Keseler et al., 2016) and RegulonDB (Santos-Zavaleta et al., 2019). Now, using the sequencing methods explained above we were able to identify candidate binding sites. For a number of cases, these putative binding sites were then used to synthesize oligonucleotide probes to enrich and identify their corresponding putative transcription factor using mass spectrometry. While Figure 4.4 gives a sense of the kinds of regulatory architectures we discovered in this study, our entire collection of regulatory cartoons can be found in Figure 4.9.

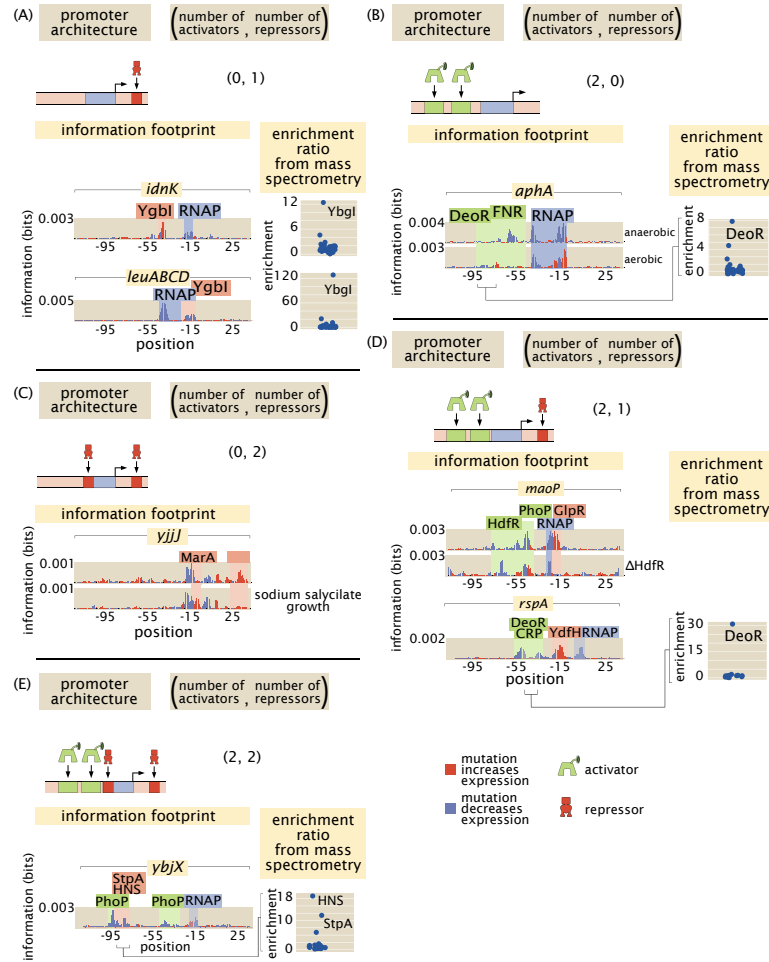


Figure 4.4: Newly discovered or updated regulatory architectures. Examples of information footprints, gene knockouts, and mass spectrometry data used to identify transcription factors for five genes. (A) Examples of simple repression, i.e. (0, 1) architectures where the locations of the putative binding sites are highlighted in red and the identities of the bound transcription factors are revealed in the mass spectrometry data. (B) An example of a (2, 0) architecture. During aerobic growth FNR is inactive, but the DeoR site now has a significant effect on expression. (C) An example of a (0, 2) architecture. *yjiJ* is regulated by MarA, which is only active in growth with sodium salicylate, and an unknown repressor. (D) An example of a (2, 1) architecture. (E) An example of a (2, 2) architecture.

A recent paper christened that part of the *E. coli* genome for which the function of the genes is unknown the y-ome (Ghatak et al., 2019). Their surprising finding is that roughly 35% of the genes in the *E. coli* genome are functionally unannotated. The situation is likely worse for other organisms. For many of the genes in the y-ome, we remain similarly ignorant of how those genes are regulated. Figures 4.4

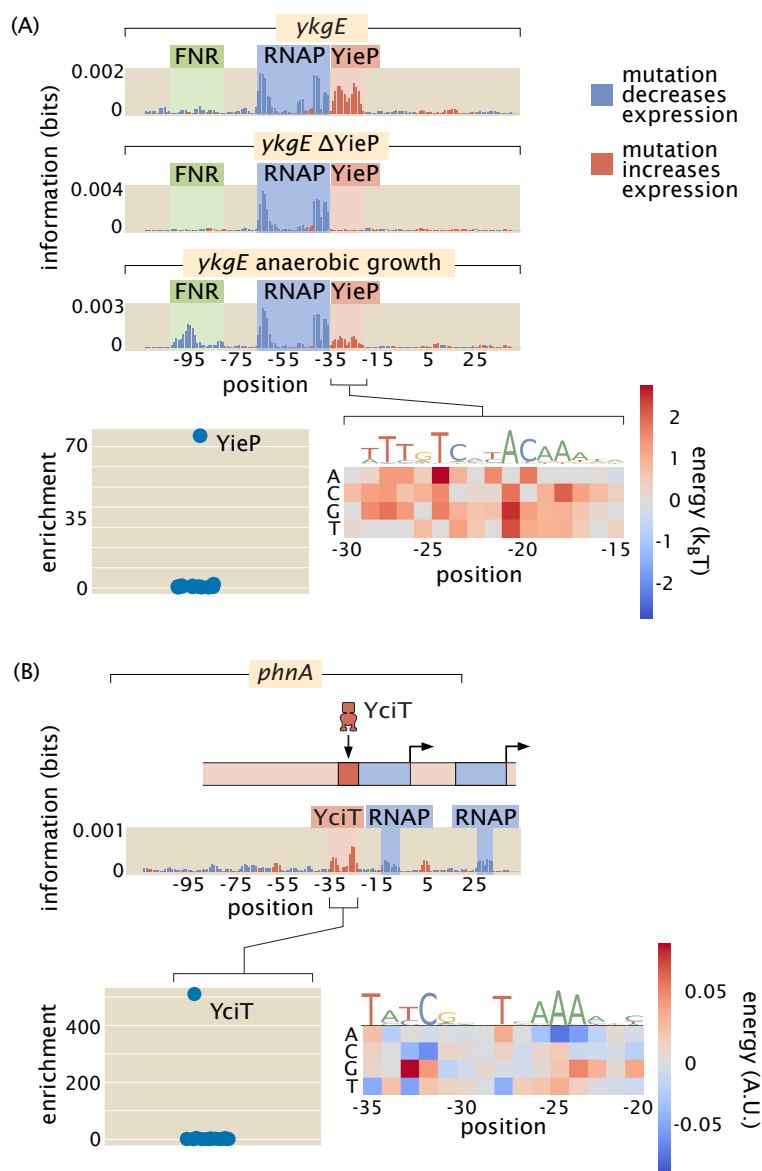


Figure 4.5: Examples of the insight gained by Reg-Seq in the context of promoters with no previously known regulatory information. (A) From the information footprint of the *ykgE* promoter under different growth conditions, we can identify a repressor binding site downstream of the RNAP binding site. From the enrichment of proteins bound to the DNA sequence of the putative repressor as compared to a control sequence, we can identify YieP as the transcription factor bound to this site as it has a much higher enrichment ratio than any other protein. Lastly, the binding energy matrix for the repressor site along with corresponding sequence logo shows that the wild type sequence is the strongest possible binder and it displays an imperfect inverted repeat symmetry. (B) Illustration of a comparable dissection for the *phnA* promoter.

and 4.5 provide several examples from the y-ome, of genes and transcription factors for which little to nothing was previously known. As shown in Figure 4.5, our study has found the first examples that we are aware of in the entire *E. coli* genome of a binding site for YciT. These examples are intended to show the outcome of the methods developed here and to serve as an invitation to browse the online resource (www.rpgroup.caltech.edu/RNAseq_SortSeq/interactive_a) to see many examples of the regulation of y-ome genes.

The ability to find binding sites for both widely acting regulators and transcription factors which may have only a few sites in the whole genome allows us to get an in-depth and quantitative view of any given promoter. As indicated in Figures 4.5(A) and (B), we were able to perform the relevant search and capture for the transcription factors that bind our putative binding sites. In both of these cases, we now hypothesize that these newly discovered binding site-transcription factor pairs exert their control through repression. The ability to extract the quantitative features of regulatory control through energy matrices means that we can take a nearly unstudied gene such as *ykgE*, which is regulated by an understudied transcription factor YieP, and quickly get to the point at which we can do quantitative modeling in the style that we and many others have performed on the *lac* operon (Vilar and Leibler, 2003; Bintu et al., 2005; Kinney, Murugan, et al., 2010; Garcia and Phillips, 2011; Vilar and Saiz, 2013; Barnes et al., 2019; Phillips et al., 2019).

One of the revealing case studies that demonstrates the broad reach of our approach for discovering regulatory architectures is offered by the insights we have gained into two widely acting regulators, GlpR (Schweizer, Boos, and Larson, 1985) and FNR (Körner, Sofia, and Zumft, 2003; Kargeti and Venkatesh, 2017). In both cases, we have expanded the array of promoters that they are now known to regulate. Further, these two case studies illustrate that even for widely acting transcription factors, there is a large gap in regulatory knowledge and the approach advanced here has the power to discover new regulatory motifs. The newly discovered binding sites in Figure 4.6(A) more than double the number of operons known to be regulated by GlpR as reported in RegulonDB (Santos-Zavaleta et al., 2019). We found 5 newly regulated operons in our data set, even though we were not specifically targeting GlpR regulation. Although the number of example promoters across the genome that we considered is too small to make good estimates, finding 5 regulated operons

out of approximately 100 examined operons supports the claim that GlpR widely regulates and many more of its sites would be found in a full search of the genome. The regulatory roles revealed in Figure 4.6(A) also reinforce the evidence that GlpR is a repressor.

For the GlpR-regulated operons newly discovered here, we found that this repressor binds strongly in the presence of glucose while all other growth conditions result in greatly diminished, but not entirely abolished, binding (Figure 4.6(A)). As there is no previously known direct molecular interaction between GlpR and glucose and the repression is reduced but not eliminated, the derepression in the absence of glucose is likely an indirect effect. As a potential mechanism of the indirect effect, *gpsA* is known to be activated by CRP (Seoh and Tai, 1999), and GpsA is involved in the synthesis of glycerol-3-phosphate (G3P), a known binding partner of GlpR which disables its repressive activity (Larsons et al., 1987). Thus in the presence of glucose GpsA and consequently G3P will be found in low concentration, ultimately allowing GlpR fulfill its role as a repressor.

Prior to this study, there were 4 operons known to be regulated by GlpR, each with between 4 and 8 GlpR binding sites (Gama-Castro et al., 2016), where the absence of glucose and the partial induction of GlpR was not enough to prompt a notable change in gene expression (Lin, 1976). These previously explored operons seemingly are regulated as part of an AND gate, where high G3P concentration *and* an absence of glucose is required for high gene expression. By way of contrast, we have discovered operons whose regulation appears to be mediated by a single GlpR site per operon. With only a single site, GlpR functions as an indirect glucose sensor, as only the absence of glucose is needed to relieve repression by GlpR.

The second widely acting regulator our study revealed, FNR, has 151 binding sites already reported in RegulonDB and is well studied compared to most transcription factors (Gama-Castro et al., 2016). However, the newly discovered FNR sites displayed in Figure 4.6(B) demonstrate that even for well-understood transcription factors there is much still to be uncovered. Our information footprints are in agreement with previous studies suggesting that FNR acts as an activator. In the presence of O₂, dimeric FNR is converted to a monomeric form and its ability to bind DNA is greatly reduced (Myers et al., 2013). Only in low oxygen conditions did we

observe a binding signature from FNR, and we show a representative example of the information footprint from one of 11 growth conditions with plentiful oxygen in Figure 4.6(B).

We observe quantitatively how FNR affects the expression of *fdhE* both directly through transcription factor binding (Figure 4.7(A)) and indirectly through increased expression of ArcA (Figure 4.7(B)). Also, fully understanding even a single operon often requires investigating several regulatory regions as we have in the case of *fdoGHI-fdhE* by investigating the main promoter for the operon as well as the promoter upstream of *fdhE*. 36% of all multi-gene operons have at least one TSS which transcribes only a subset of the genes in the operon (Conway et al., 2014). Regulation within an operon is even more poorly studied than regulation in general. The main promoter for *fdoGHI-fdhE* has a repressor binding site, which demonstrates that there is regulatory control of the entire operon. However, we also see in Figure 4.7(B) that there is control at the promoter level, as *fdhE* is regulated by both ArcA and FNR and will therefore be upregulated in anaerobic conditions (Compan and Touati, 1994). The main TSS transcribes all four genes in the operon, while the secondary site shown in Figure 4.7(B) only transcribes *fdhE*, and therefore anaerobic conditions will change the stoichiometry of the proteins produced by the operon. At the higher throughput that we use in this experiment it becomes feasible to target multiple promoters within an operon as we have done with *fdoGHI-fdhE*. We can then determine under what conditions an operon is internally regulated. Figure 4.7 also makes it clear that for cases such as *fdoGHI-fdhE*, there are many subtleties both in the interpretation of the information footprints and in the construction of regulatory cartoons that are simultaneously accurate and transparent. A crucial next step in the development of these analyses is to move from manual curation of the data to automated statistical analyses that can help make sense of these complicated datasets.

By examining the over 100 promoters considered here, grown under 12 growth conditions, we have a total of more than 1000 information footprints and data sets. In this age of big data, methods to explore and draw insights from that data are crucial. To that end, as introduced in Figure 4.8, we have developed an online resource (see www.rpgroup.caltech.edu/RNAseq_SortSeq/interactive_a) that makes it possible for anyone who is interested to view our data and draw their own biological conclusions. Information footprints for any combination of gene and

growth condition are displayed via drop down menus. Each identified transcription factor binding site or transcription start site is marked, and energy matrices for all transcription factor binding sites are displayed. In addition, for each gene, we feature a simple cartoon-level schematic that captures our now current best understanding of the regulatory architecture and resulting mechanism.

The interactive figure in question was invaluable in identifying transcription factors, such as GlpR, whose binding properties vary depending on growth condition. As sigma factor availability also varies greatly depending on growth condition, studying the interactive figure identified many of the secondary RNAP sites present. The interactive figure provides a valuable resource both to those who are interested in the regulation of a particular gene and those who wish to look for patterns in gene regulation across multiple genes or across different growth conditions.

Regulatory cartoons

The final coarse grained output from Reg-Seq can be represented as regulatory cartoons in Fig 4.9, which display the transcription factor binding sites and TF identities if known. Binding site locations are coarse grained, along with RNAP sites (in blue). In other words, if a repressor binding site is displayed overlapping the RNAP site, then the repressor overlaps the RNAP site, and if the repressor is displayed downstream of the RNAP site, then it is downstream of the RNAP site. However, just the distance that the repressor is displayed downstream of the RNAP site does not directly reflect the number of base pairs downstream that the repressor binding site is from the RNAP.

4.3 Discussion

The study of gene regulation is one of the centerpieces of modern biology. As a result, it is surprising that in the genome era, our ignorance of the regulatory landscape in even the best-understood model organisms remains so vast. Despite understanding the regulation of transcription initiation in bacterial promoters (Browning and Busby, 2016), and how to tune their expression, we lack an experimental framework to unravel understudied promoter architectures at scale. As such, in our view one of the grand challenges of the genome era is the need to uncover the regulatory landscape for each and every organism with a known genome sequence. Given the ability to read and write DNA sequence at will, we are convinced that to make

that reading of DNA sequence truly informative about biological function and to give that writing the full power and poetry of what Crick christened “the two great polymer languages”, we need a full accounting of how the genes of a given organism are regulated and how environmental signals communicate with the transcription factors that mediate that regulation - the so-called “allosterome” problem (Lindsley and Rutter, 2006). The work presented here provides a general methodology for making progress on the former problem and also demonstrates that, by performing Reg-Seq in different growth conditions, we can make headway on the latter problem as well.

The advent of cheap DNA sequencing offers the promise of beginning to achieve that grand challenge goal in the form of MPRA reviewed in Kinney and McCandlish, 2019. A particular implementation of such methods was christened Sort-Seq (Kinney, Murugan, et al., 2010) and was demonstrated in the context of well understood regulatory architectures. A second generation of the Sort-Seq method (Belliveau et al., 2018) established experiments through the use of DNA-affinity chromatography and mass spectrometry which made it possible to identify the transcription factors that bind the putative binding sites discovered by Sort-Seq. But there were critical shortcomings in the method, not least of which was that it lacked the scalability to uncover the regulatory genome on a genome-wide basis.

The work presented here builds on the foundations laid in the previous studies by invoking RNA-Seq as a readout of the level of expression of the promoter mutant libraries needed to infer information footprints and their corresponding energy matrices and sequence logos followed by a combination of mass spectrometry and gene knockouts to identify the transcription factors that bind those sites. The case studies described in the main text showcase the ability of the method to deliver on the promise of beginning to uncover the regulatory genome systematically. The extensive online resources hint at a way of systematically reporting those insights in a way that can be used by the community at large to develop regulatory intuition for biological function and to design novel regulatory architectures using energy matrices.

However, several shortcomings remain in the approach introduced here. First, the current implementation of Reg-Seq still largely relies on manual curation as the basis of using information footprints to generate testable regulatory hypotheses. As

described in the methods section, we have also used statistical testing as a way to convert information footprints into regulatory hypotheses, but there clearly remains much work to be done on the data analysis pipeline to improve both the power and the accuracy of this approach. In addition, these regulatory hypotheses can also be converted into gene regulatory models using statistical physics (Buchler, Gerland, and Hwa, 2003; Bintu et al., 2005). However, here too, as the complexity of the regulatory architectures increases, it will be of great interest to use automated model generation as suggested in a recent biophysically-based neural network approach (Tareen and Kinney, 2019).

A second key challenge faced by the methods described here is that the mass spectrometry and the gene knockout confirmation aspects of the experimental pipeline remain low-throughput. To overcome this, we have begun to explore a new generation of experiments such as *in vitro* binding assays that will make it possible to accomplish transcription factor identification at higher throughput. Specifically, we are exploring multiplexed mass spectrometry measurements and multiplexed Reg-Seq on libraries of gene knockouts as ways to break the identification bottleneck.

Another shortcoming of the current implementation of the method is that it would miss regulatory action at a distance. Indeed, our laboratory has invested a significant effort in exploring such long-distance regulatory action in the form of DNA looping in bacteria and VDJ recombination in jawed vertebrates. It is well known that transcriptional control through enhancers in eukaryotic regulation is central in contexts ranging from embryonic development to hematopoiesis (Melnikov et al., 2012). The current incarnation of the methods described here have focused on contiguous regions in the vicinity of the transcription start site. Clearly, to go further in dissecting the entire regulatory genome, these methods will have to be extended to non-contiguous regions of the genome.

The findings from this study provide a foundation for systematically performing genome-wide regulatory dissections. We have developed a method to pass from complete regulatory ignorance to designable regulatory architectures and we are hopeful that others will adopt these methods with the ambition of uncovering the regulatory architectures that preside over their organisms of interest.

4.4 Methods

Library construction

Promoter variants were synthesized on a microarray (TWIST Bioscience, San Francisco, CA). The sequences were designed computationally such that each base in the 160 bp promoter region has a 10% probability of being mutated. For each given promoter's library, we ensured that the mutation rate as averaged across all sequences was kept between 9.5% and 10.5%, otherwise the library was regenerated. There are an average of 2200 unique promoter sequences per gene (for an analysis of how our results depend upon number of unique promoter sequences see Figure B.5). An average of 5 unique 20 base pair barcodes per variant promoter was used for the purpose of counting transcripts. The barcode was inserted 110 base pairs from the 5' end of the mRNA, containing 45 base pairs from the targeted regulatory region, 64 base pairs containing primer sites used in the construction of the plasmid, and 11 base pairs containing a three frame stop codon. All the sequences are listed in Supplementary Table 1. Following the barcode there is an RBS and a GFP coding region. Mutated promoters were PCR amplified and inserted by Gibson assembly into the plasmid backbone of pJK14 (SC101 origin) (Kinney, Murugan, et al., 2010). Constructs were electroporated into *E. coli* K-12 MG1655 (Blattner, 1997).

RNA preparation and sequencing

Cells were grown to an optical density of 0.3 and RNA was then stabilized using Qiagen RNA Protect (Qiagen, Hilden, Germany). Lysis was performed using lysozyme (Sigma Aldrich, Saint Louis, MO) and RNA was isolated using the Qiagen RNA Mini Kit. Reverse transcription was performed using Superscript IV (Invitrogen, Carlsbad, CA) and a specific primer for the labeled mRNA. qPCR was performed to check the level of DNA contamination and the mRNA tags were PCR amplified and Illumina sequenced. Within a single growth condition, all promoter variants for all regulatory regions were tested in a single multiplexed RNA-Seq experiment. All sequencing was carried out by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech (HiSeq 2500) on a 100 bp single read flow cell or using the sequencing services from NGX Bio on a 250 bp or 150 base paired end flow cell.

Analysis of sequencing results

To determine putative transcription factor binding sites, we first compute the effect of mutations on gene expression at a base pair-by-base pair level using information footprints. The information footprints are a hypothesis generating tool and we choose which regions to further investigate using techniques such as mass spectrometry by visually inspecting the data for regions of 10 to 20 base pairs that have high information content compared to background. Our technique currently relies on using human intuition to determine binding sites, but to validate these choices and to capture all regions important for gene expression we computationally identify regions where gene expression is changed significantly up or down by mutation ($p < 0.01$), and discard any potential sites which do not fit this criteria. We infer the effect of mutation using Markov Chain Monte Carlo, and we use the distribution of parameters from the inference to form a 99 % confidence interval for the average effect of mutation across a 15 base pair region. We include binding sites that are statistically significant at the 0.01 level in any of the tested growth conditions.

Many false positives will be secondary RNAP sites and we remove from consideration any sites that resemble RNAP sites. We fit energy matrices to each of the possible binding sites and use the preferred DNA sequence for binding to identify the RNAP sites. We use both visual inspection to compare the preferred sequence to known consensus sequences for each of the *E. coli* sigma factor binding sites (for example, do the preferred bases in the energy matrix have few mismatches to the TGNTATAAT extended minus 10 for σ^{70} sites), and the TOMTOM tool (Gupta et al., 2007) to computationally compare the potential site to examples of σ^{70} , σ^{38} , and σ^{54} sites that we determined in this experiment. For further details see Figure B.6. We discard any sites that have a p-value of similarity with an RNAP site of less than 5×10^{-3} in the TOMTOM analysis or are deemed to be too visually similar to RNAP sites. If a single site contains an RNAP site along with a transcription factor site we remove only those bases containing the probable RNAP site. This results in 95 identified transcription factor binding regions.

For primary RNAP sites, we include a list of probable sigma factor identities as Supplementary Table 2 Sites are judged by visual similarity to consensus binding sites. Those sites where the true sigma factor is unclear due to overlapping binding sites are omitted. Overlapping binding sites (from multiple TFs or RNAP sites) in

general can pose issues for this method. In many cases, looking at growth conditions where only one of the relevant transcription factors is present or active is an effective way to establish site boundaries and infer correct energy matrices. For sites where no adequate growth condition can be found, or when a TF overlaps with an RNAP site, the energy matrix will not be reflective of the true DNA-protein interaction energies. If the TFs in overlapping sites are composed of one activator and one repressor, then we use the point at which the effect of mutation shifts from activator-like to repressor-like as a demarcation point between binding sites. We see a case of a potentially overlooked repressor due to overlapping sites in Figure 4.4(B), as there are several repressor like bases overlapping the RNAP -10 site and the effect weakens in low oxygen growth. However, due to the effect of the RNAP site, when averaged over a potential 15 base pair region, the repressor-like bases do not have a significant effect on gene expression.

DNA affinity chromatography and mass spectrometry

Upon identifying a putative transcription factor binding site, we used DNA affinity chromatography, as done in (Belliveau et al., 2018) to isolate and enrich for the transcription factor of interest. In brief, we order biotinylated oligos of our binding site of interest (Integrated DNA Technologies, Coralville, IA) along with a control, "scrambled" sequence, that we expect to have no specificity for the given transcription factor. We tether these oligos to magnetic streptavidin beads (Dynabeads MyOne T1; ThermoFisher, Waltham, MA), and incubate them overnight with whole cell lysate grown in the presences of either heavy (with ^{15}N) or light (with ^{14}N) lysine for the experimental and control sequences, respectively. The next day, proteins are recovered by digesting the DNA with the PstI restriction enzyme (New England Biolabs, Ipswich, MA), whose cut site was incorporated into all designed oligos.

Protein samples were then prepared for mass spectrometry by either in-gel or in-solution digestion using the Lys-C protease (Wako Chemicals, Osaka, Japan). Liquid chromatography coupled mass spectrometry (LC-MS) was performed as previously described by (Belliveau et al., 2018), and is further discussed in the A. SILAC labeling was performed by growing cells (Δ LysA) in either heavy isotope form of lysine or its natural form.

It is also important to note that while we relied on the SILAC method to identify the TF identity for each promoter, our approach doesn't require this specific technique. Specifically, our method only requires a way to contrast between the copy number of proteins bound to a target promoter in relation to a scrambled version of the promoter. In principle, one could use multiplexed proteomics based on isobaric mass tags (Pappireddi, Martin, and Wüehr, 2019) to characterize up to 10 promoters in parallel. Isobaric tags are reagents used to covalently modify peptides by using the heavy-isotope distribution in the tag to encode different conditions. The most widely adopted methods for isobaric tagging are the isobaric tag for relative and absolute quantitation (iTRAQ) and the tandem mass tag (TMT). This multiplexed approach involves the fragmentation of peptide ions by colliding with an inert gas. The resulting ions are resolved in a second MS-MS scan (MS2).

Only a subset (13) of all transcription factor targets were identified by mass spectrometry due to limitations in scaling the technique to large numbers of targets. The transcription factors identified by this method are enriched more than any other DNA binding protein, with $p < 0.01$ using the outlier detection method as outlined by Cox and Mann, 2008, with corrections for multiple hypothesis testing using the method proposed by Benjamini and Hochberg, 1995.

Construction of knockout strains

Conducting DNA affinity chromatography followed by mass spectrometry on putative binding sites resulted in potential candidates for the transcription factors that are responsible for the information contained at a given promoter region. For some cases, to verify that a given transcription factor is, in fact, regulating a given promoter, we repeated the RNA sequencing experiments on strains with the transcription factor of interest knocked out.

To construct the knockout strains, we ordered strains from the Keio collection (Yamamoto et al., 2009) from the Coli Genetic Stock Center. These knockouts were put in a MG1655 background via phage P1 transduction and verified with Sanger sequencing. To remove the kanamycin resistance that comes with the strains from the Keio collection, we transformed in the pCP20 plasmid, induced FLP recombinase, and then selected for colonies that no longer grew on either kanamycin or ampi-

cillin. Finally, we transformed our desired promoter libraries into the constructed knockout strains, allowing us to perform the RNA sequencing in the same context as the original experiments.

Code and Data Availability

All code used for processing data and plotting as well as the final processed data, plasmid sequences, and primer sequences can be found on the GitHub repository (https://github.com/RPGroup-PBoC/RNAseq_SortSeq) doi:10.5281/zenodo.3628117.

Energy matrices were generated using the MPATHIC software (Ireland and Kinney, 2016). All raw sequencing data is available at the Sequence Read Archive (accession no. PRJNA599253 and PRJNA603368). All inferred information footprints and energy matrices can be found on the CalTech data repository doi:10.22002/D1.1331.

All mass spectrometry raw data is available on the CalTech data repository doi:10.22002/d1.1336

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
(0, 0)	<i>acuI</i>	0	0	0	
	<i>adiY</i>	0	0	0	
	<i>arcB</i>	0	0	0	
	<i>coaA</i>	0	0	0	
	<i>dnaE</i>	0	0	0	
	<i>ecnB</i>	0	0	0	
	<i>holC</i>	0	0	0	
	<i>hslU</i>	0	0	0	
	<i>htrB</i>	0	0	0	
	<i>modE</i>	0	0	0	
	<i>motAB-cheAW</i>	0	0	0	
	<i>poxB</i>	0	0	0	
	<i>rcsF</i>	0	0	0	
	<i>rumB</i>	0	0	0	
	<i>sbcB</i>	0	0	0	
	<i>sdaB</i>	0	0	0	
	<i>ybdG</i>	0	0	0	
	<i>ybiP</i>	0	0	0	
	<i>ybjL</i>	0	0	0	
	<i>ybjT</i>	0	0	0	
	<i>yehS</i>	0	0	0	
	<i>yehT</i>	0	0	0	
	<i>yfhG</i>	0	0	0	
	<i>ygdH</i>	0	0	0	
	<i>ygeR</i>	0	0	0	
	<i>yggW</i>	0	0	0	
	<i>ygiP</i>	0	0	0	
	<i>ynaI</i>	0	0	0	
	<i>yqhC</i>	0	0	0	
	<i>zapB</i>	0	0	0	
	<i>zupT</i>	0	0	0	
	<i>amiC</i>	0	0	0	
(0, 1)	<i>aegA</i>	1	0	0	
					Known binding location (NsrR) (Partridge et al., 2009)
	<i>bdcR</i>	1	0	1	
	<i>dicC</i>	0	1	0	
	<i>fdoH</i>	1	0	0	
	<i>groSL</i>	1	0	0	

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
	<i>idnK</i>	1	0	1	Mass-Spectrometry (YgbI)
	<i>leuABCD</i>	1	0	1	Mass-Spectrometry (YgbI)
	<i>pcm</i>	1	0	0	
	<i>yedK</i>	1	0	1	Mass-Spectrometry (TreR)
	<i>rapA</i>	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR)
	<i>sdiA</i>	1	0	0	
	<i>tar</i>	1	0	0	
	<i>tff-rpsB-tsrf</i>	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (GlpR)
	<i>thiM</i>	1	0	0	
	<i>tig</i>	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (GlpR)
	<i>ycgB</i>	1	0	0	
	<i>ydjA</i>	1	0	0	
	<i>yedJ</i>	1	0	0	
	<i>ycbZ</i>	1	0	0	
	<i>phnA</i>	1	0	1	Mass-Spectrometry (YciT)
	<i>mutM</i>	1	0	0	
	<i>rhlE</i>	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Mass-Spectrometry (GlpR)
	<i>uvrD</i>	1	0	1	Bioinformatic (LexA)
	<i>dusC</i>	1	0	0	
	<i>ftsK</i>	0	1	0	
	<i>znuA</i>	0	1	0	
(1, 0)	<i>waaA-coaD</i>	1	0	0	
	<i>cra</i>	1	0	0	
	<i>iap</i>	1	0	0	
	<i>araC</i>	0	1	0	
	<i>minC</i>	1	0	0	

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
	<i>ybeZ</i>	1	0	0	
	<i>mscM</i>	1	0	0	
	<i>mscS</i>	1	0	0	
	<i>rlmA</i>	1	0	0	
	<i>thrLABC</i>	1	0	0	
	<i>yeyQ</i>	1	0	1	Growth condition Knockout (FNR), Bioinformatic (FNR)
	<i>dgoR</i>	0	1	0	Mass- Spectrometry (DgoR)
	<i>lac</i>	0	1	0	Mass- Spectrometry (LacI)
(0, 2)	<i>yecE</i>	2	0	1	Mass- Spectrometry (HU)
	<i>yjjJ</i>	2	0	1	Growth condition Knockout (MarA), Bioinformatic (MarA)
	<i>dcm</i>	2	0	1	Mass- Spectrometry (HNS)
	<i>marR</i>	0	2	0	Mass- Spectrometry (MarR)
(1, 1)	<i>ilvC</i>	2	0	1	Mass- Spectrometry (IlvY)
	<i>ybiO</i>	2	0	0	
	<i>yehU</i>	2	0	1	Growth condition Knockout (FNR), Bioinformatic (FNR)
	<i>ykgE</i>	2	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass- Spectrometry(YieP) Knockout (YieP)

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
	<i>ymgG</i>	2	0	0	
	<i>znuCB</i>	1	1	0	
(2, 0)	<i>aphA</i>	2	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass-Spectrometry (DeoR)
	<i>arcA</i>	2	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass-Spectrometry (FNR, CpxR)
	<i>asnA</i>	2	0	0	
	<i>fdhE</i>	2	0	2	Growth condition Knockout (FNR, ArcA), Bioinformatic (FNR, ArcA), Knockout (ArcA)
	<i>xylF</i>	0	2	0	
	<i>mscL</i>	2	0	0	
(2, 1)	<i>maoP</i>	3	0	3	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (PhoP, HdfR, GlpR)
	<i>rspA</i>	1	2	1	Mass-Spectrometry (DeoR)
(1, 2)	<i>dinJ</i>	3	0	0	
(2, 2)	<i>ybjX</i>	4	0	4	Bioinformatic (PhoP), Mass-Spectrometry (HNS, StpA)
(3, 0)	<i>araAB</i>	0	3	0	
	<i>xylA</i>	0	3	0	
	<i>yicI</i>	3	0	0	
(0, 4)	<i>relBE</i>	0	4	0	Mass-Spectrometry (RelBE)

Table 4.2: All genes investigated in this study categorized according to their regulatory architecture

The architecture is given as (number of activators, number of repressors). The table also lists the number of newly discovered binding sites, previously known binding sites, and number of identified transcription factors. The evidence used for the transcription factor identification is given in the final column.

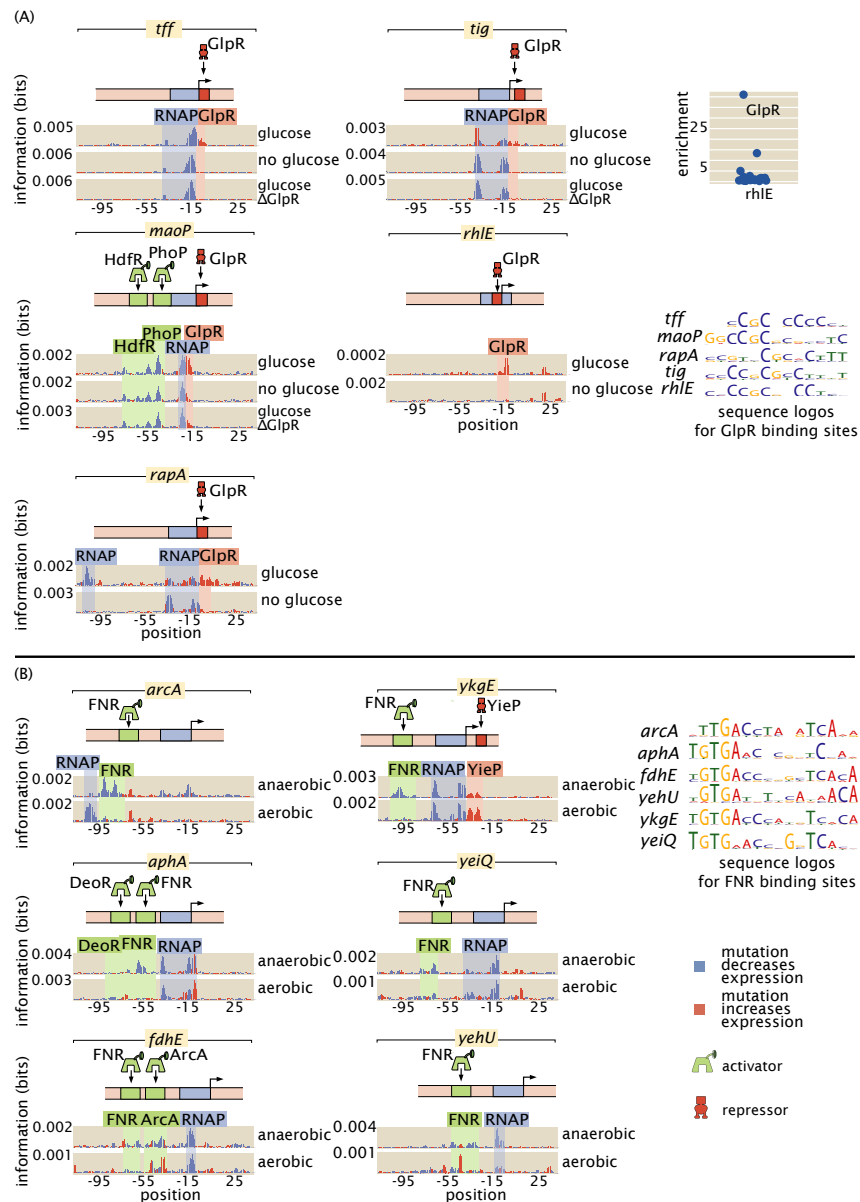


Figure 4.6: Reg-Seq analysis of broadly-acting transcription factors. (A) GlpR as a widely-acting regulator. Here we show the many promoters which we found to be regulated by GlpR, all of which were previously unknown. GlpR was demonstrated to bind to *rhIE* by mass spectrometry enrichment experiments as shown in the top right. Binding sites in the *tff*, *tig*, *maoP*, *rhIE*, and *rapA* have similar DNA binding preferences as seen in the sequence logos and each TF binding site binds strongly only in the presence of glucose. These similarities suggest that the same TF binds to each site. To test this hypothesis we knocked out GlpR and ran the Reg-Seq experiments for *tff*, *tig*, and *maoP*. We see that knocking out GlpR removes the binding signature of the TF. (B) FNR as a global regulator. FNR is known to be upregulated in anaerobic growth, and here we found it to regulate a suite of six genes. In growth conditions with prevalent oxygen the putative FNR sites are weakened, and the DNA binding preference of the six sites are shown to be similar from the sequence logos displayed on the right.

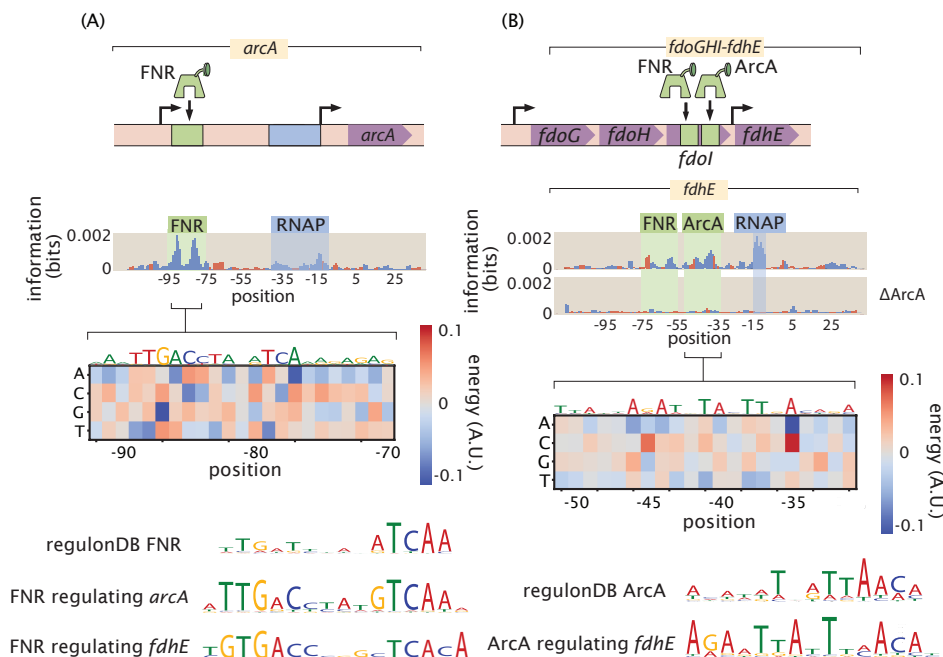


Figure 4.7: Inspection of an anaerobic respiration genetic circuit. (A) Here we see not only how the *arcA* promoter is regulated, but also the role this transcription factor plays in the regulation of another promoter. (B) Intra-operon regulation of *fdhE* by both FNR and ArcA. A TOMTOM (Gupta et al., 2007) search of the binding motif found that ArcA was the most likely candidate for the transcription factor. A knockout of ArcA demonstrates that the binding signature of the site, and its associated RNAP site, are no longer significant determinants of gene expression.

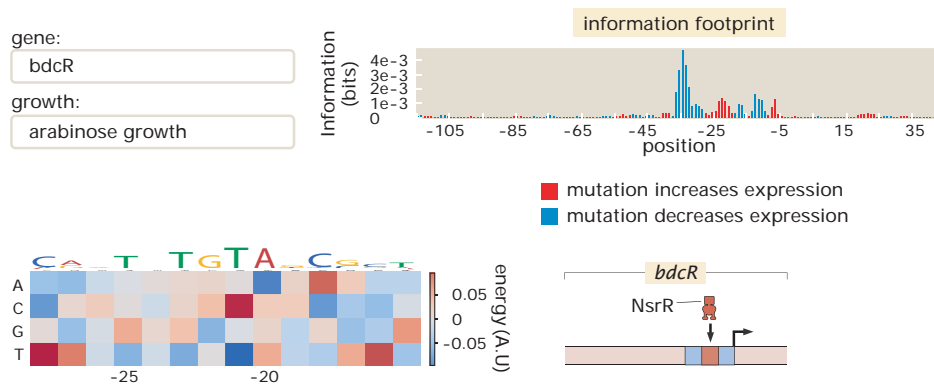


Figure 4.8: Representative view of the interactive figure that is available online. This interactive figure captures the entirety of our dataset. Each figure features a drop-down menu of genes and growth conditions. For each such gene and growth condition, there is a corresponding information footprint revealing putative binding sites, an energy matrix that shows the strength of binding of the relevant transcription factor to those binding sites and a cartoon that schematizes the newly-discovered regulatory architecture of that gene.

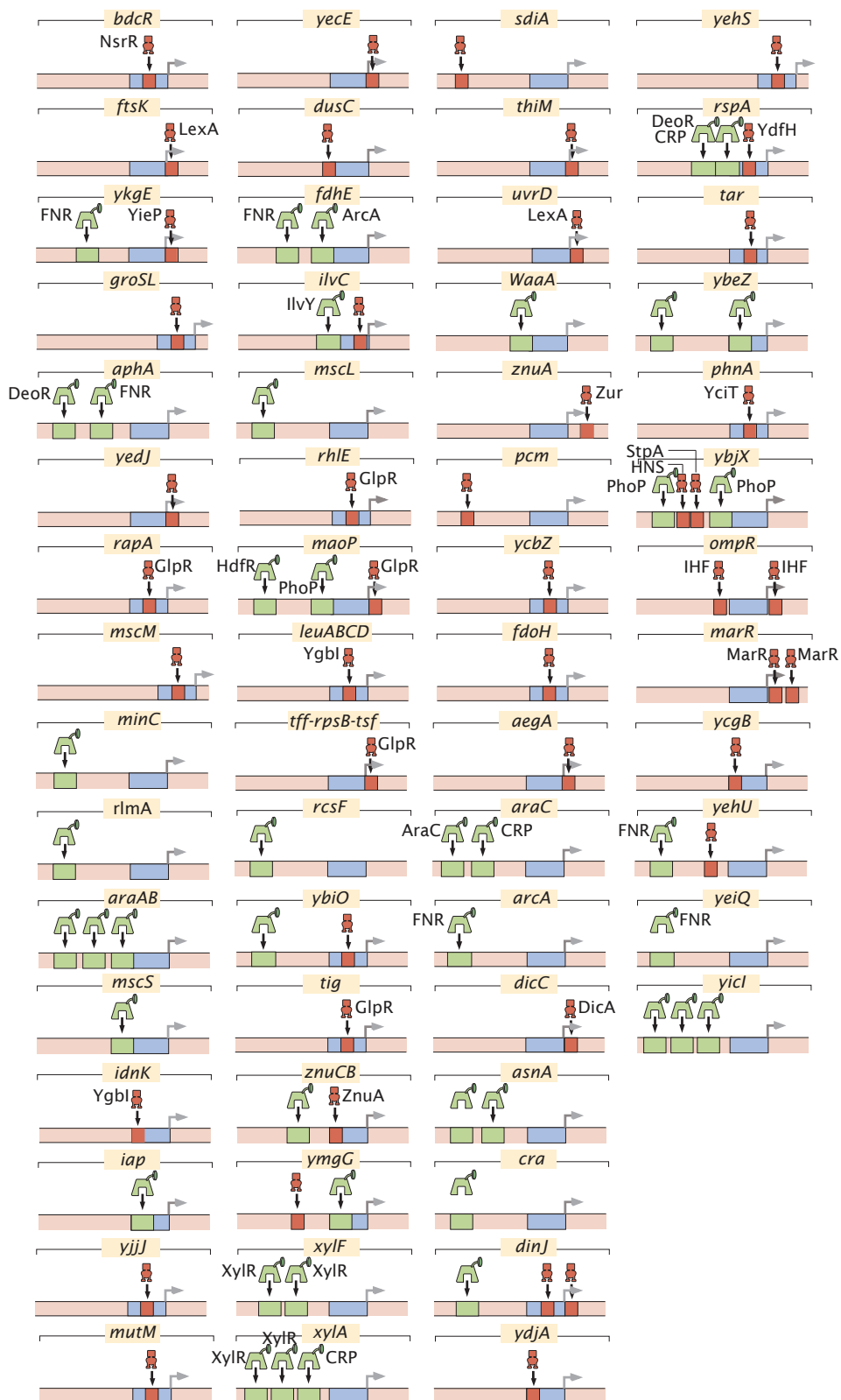


Figure 4.9: All regulatory cartoons for genes considered in Reg-Seq.

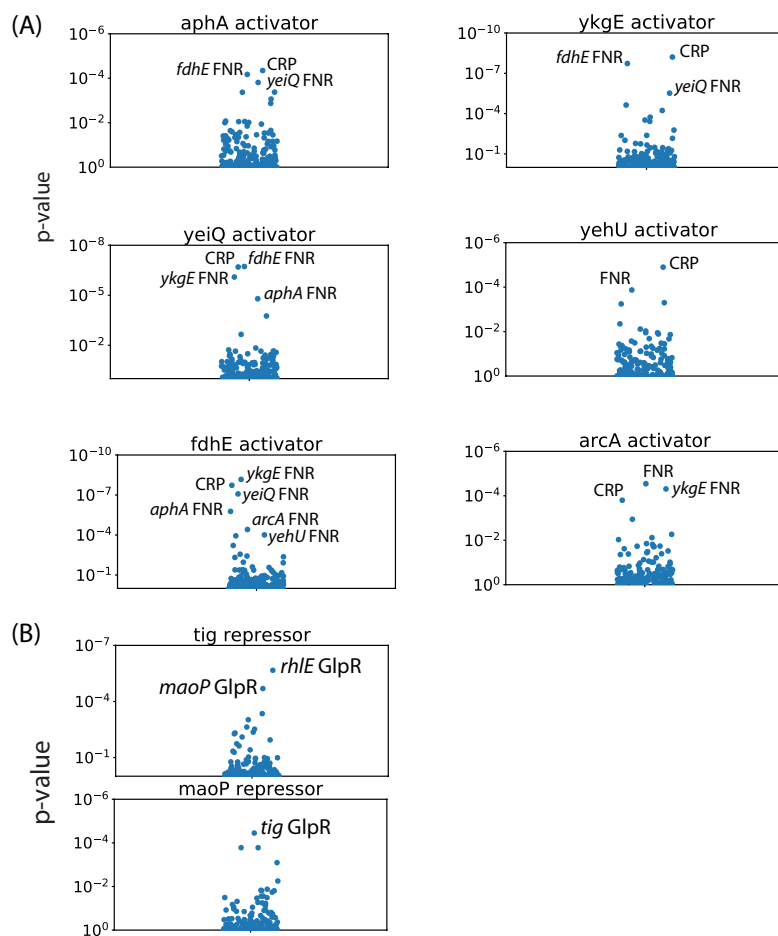


Figure 4.10: The p-value distribution from TOMTOM for comparisons to FNR and GlpR binding sites. (A) TOMTOM comparisons for the FNR sites for *aphA*, *ykgE*, *yeiQ*, *yehU*, *fdhE*, *arcA*. For each case the site is compared to all other discovered sites from Reg-Seq as well as all binding site motifs derived from RegulonDB. (B) TOMTOM comparisons for the GlpR sites for *tig* and *maoP*. For each case the site is compared to all other discovered sites from Reg-Seq as well as all binding site motifs derived from RegulonDB.

BIBLIOGRAPHY

- Barnes, Stephanie L. et al. (2019). “Mapping DNA sequence to transcription factor binding energy *in vivo*”. In: *PLoS Computational Biology* 15.2, pp. 1–29. doi: 10.1371/journal.pcbi.1006226.
- Belliveau, Nathan M. et al. (2018). “Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.21, E4796–E4805. doi: 10.1073/pnas.1722055115.
- Benjamini, Yoav and Yosef Hochberg (Jan. 1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.
- Berg, O. G. and P. H. von Hippel (1987). “Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters”. In: *J Mol Biol* 193.4, pp. 723–50. doi: 10.1016/0022-2836(87)90354-8.
- Bintu, Lacramioara et al. (Apr. 2005). “Transcriptional regulation by the numbers: models”. en. In: *Current Opinion in Genetics & Development*. Chromosomes and expression mechanisms 15.2, pp. 116–124. doi: 10.1016/j.gde.2005.02.007.
- Blattner, F. R. (Sept. 1997). “The Complete Genome Sequence of *Escherichia coli* K-12”. en. In: *Science* 277.5331, pp. 1453–1462. doi: 10.1126/science.277.5331.1453.
- Browning, Douglas F and Busby (2016). “Local and global regulation of transcription initiation in bacteria”. In: *Nature Reviews Microbiology*, pp. 638–650. doi: 10.1038/nrmicro.2016.103.
- Buchler, Nicolas E, Ulrich Gerland, and Terence Hwa (Apr. 2003). “On schemes of combinatorial transcription logic.” In: *Proceedings of the National Academy of Sciences* 100.9, pp. 5136–5141. doi: 10.1073/pnas.0930314100.
- Chure, Griffin et al. (2019). “Predictive shifts in free energy couple mutations to their phenotypic consequences”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.37, pp. 18275–18284. doi: 10.1073/pnas.1907869116.
- Compan, Inès and Danlèle Touati (1994). “Anaerobic activation of *arcA* transcription in *Escherichia coli*: roles of Fnr and ArcA”. en. In: *Molecular Microbiology* 11.5, pp. 955–964. doi: 10.1111/j.1365-2958.1994.tb00374.x.
- Conway, Tyrrell et al. (July 2014). “Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing”. en. In: *mBio* 5.4. Ed. by Sankar Adhya, e01442–14. doi: 10.1128/mBio.01442-14.

- Cox, Jürgen and Mann (Dec. 2008). “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification”. en. In: *Nature Biotechnology* 26.12, pp. 1367–1372. doi: 10.1038/nbt.1511.
- Fulco et al. (2019). “Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations”. In: *Nature Genetics* 51.12, pp. 1664–1669. doi: 10.1038/s41588-019-0538-0.
- Gama-Castro, Socorro et al. (2016). “RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond”. In: *Nucleic Acids Research* 44.D1, pp. D133–D143. doi: 10.1093/nar/gkv1156.
- Gao, Ye et al. (2018). “Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655”. In: *Nucleic Acids Research* 46.20, pp. 10682–10696. doi: 10.1093/nar/gky752.
- Garcia and Phillips (July 2011). “Quantitative dissection of the simple repression input-output function”. en. In: *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–12178. doi: 10.1073/pnas.1015616108.
- Ghatak, Sankha et al. (2019). “The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function”. In: *Nucleic Acids Research* 47.5, pp. 2446–2454. doi: 10.1093/nar/gkz030.
- Goodall, Emily C. A. et al. (2018). “The Essential Genome of *Escherichia coli* K-12”. In: *mBio* 9.1. doi: 10.1128/mBio.02096-17.
- Gupta, Shobhit et al. (2007). “Quantifying similarity between motifs”. In: *Genome Biology* 8.2. doi: 10.1186/gb-2007-8-2-r24.
- Ireland, William T. and Kinney (May 2016). “MPAthic: Quantitative Modeling of Sequence-Function Relationships for massively parallel assays”. en. In: doi: 10.1101/054676.
- Jacob, Francois and Jacques Monod (1961). “On the Regulation of Gene Activity”. en. In: *Cold Spring Harbor Symposia on Quantitative Biology* 26, p. 19. doi: 10.1101/SQB.1961.026.01.024.
- Kargeti, Manika and K. V. Venkatesh (2017). “The effect of global transcriptional regulators on the anaerobic fermentative metabolism of *Escherichia coli*”. en. In: *Molecular BioSystems* 13.7, pp. 1388–1398. doi: 10.1039/C6MB00721J.
- Keseler, Ingrid M. et al. (2016). “The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12”. In: *Nucleic Acids Research* 45.D1, pp. D543–D550. doi: 10.1093/nar/gkw1003.
- Kinney and McCandlish (2019). “Massively Parallel Assays and Quantitative Sequence-Function Relationships”. In: *Annual Review of Genomics and Human Genetics* 20.1, pp. 99–127. doi: 10.1146/annurev-genom-083118-014845.

- Kinney, Anand Murugan, et al. (2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.20, pp. 9158–9163. doi: 10.1073/pnas.1004290107.
- Körner, Heinz, Heidi J. Sofia, and Walter G. Zumft (Dec. 2003). “Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs”. en. In: *FEMS Microbiology Reviews* 27.5, pp. 559–592. doi: 10.1016/S0168-6445(03)00066-4.
- Kosuri, Sriram et al. (2013). “Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.34. doi: 10.1073/pnas.1301301110.
- Kwasnieski, J. C. et al. (2012). “Complex effects of nucleotide variants in a mammalian cis-regulatory element”. In: *Proc Natl Acad Sci U S A* 109.47, pp. 19498–503. doi: 10.1073/pnas.1210678109.
- Larsons, Timothy J et al. (1987). “Purification and Characterization of the Repressor for the sn-Glycerol 3-Phosphate Regulon of *Escherichia coli* K12”. In: *Journal of Biological Chemistry* 262.33, pp. 15869–15874.
- Lin, E. C. C. (1976). “Glycerol Dissimilation and its Regulation in Bacteria”. In: *Annual Review of Microbiology* 30.1, pp. 535–578. doi: 10.1146/annurev.mi.30.100176.002535.
- Lindsley, Janet E. and Jared Rutter (2006). “Whence cometh the allosterome?” In: *Proceedings of the National Academy of Sciences of the United States of America* 103.28, pp. 10533–10535. doi: 10.1073/pnas.0604452103.
- Melnikov, Alexandre et al. (2012). “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay”. In: *Nature Biotechnology* 30.3, pp. 271–277. doi: 10.1038/nbt.2137.
- Mendoza-Vargas, Alfredo et al. (Oct. 2009). “Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*”. en. In: *PLoS ONE* 4.10. Ed. by Chad Creighton, e7526. doi: 10.1371/journal.pone.0007526.
- Mittler, G., F. Butter, and M. Mann (2009). “A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements”. In: *Genome Res* 19.2, pp. 284–93. doi: 10.1101/gr.081711.108.
- Mortazavi, Ali et al. (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5.7, pp. 621–628. doi: 10.1038/nmeth.1226.
- Myers, Kevin et al. (June 2013). “Genome-scale Analysis of *Escherichia coli* FNR Reveals Complex Features of Transcription Factor Binding”. In: *PLOS Genetics* 9.6, pp. 1–24. doi: 10.1371/journal.pgen.1003565.

- Pappireddi, Nishant, Lance Martin, and Martin Wühr (2019). “A Review on Quantitative Multiplexed Proteomics”. In: *ChemBioChem* 20.10, pp. 1210–1224. DOI: 10.1002/cbic.201800650.
- Partridge, Jonathan D. et al. (2009). “NsrR targets in the *Escherichia coli* genome: new insights into DNA sequence requirements for binding and a role for NsrR in the regulation of motility”. en. In: *Molecular Microbiology* 73.4, pp. 680–694. DOI: 10.1111/j.1365-2958.2009.06799.x.
- Patwardhan, Hiatt, et al. (2012). “Massively parallel functional dissection of mammalian enhancers *in vivo*”. In: *Nature Biotechnology* 30.3, pp. 265–70. DOI: 10.1038/nbt.2136.
- Patwardhan, C. Lee, et al. (2009). “High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis”. In: *Nature Biotechnology* 27.12, pp. 1173–1175. DOI: 10.1038/nbt.1589.
- Phillips et al. (2019). “Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression”. In: *Annual Review of Biophysics* 48, pp. 121–163.
- Razo-Mejia, Manuel et al. (2018). “Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction”. In: *Cell Systems* 6.4, 456–469.e10. DOI: 10.1016/j.cels.2018.02.004. arXiv: 1702.07460.
- Rydenfelt, M. et al. (2014). “The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*”. In: *PLoS One* 9.12, e114347. DOI: 10.1371/journal.pone.0121935.
- Santos-Zavaleta, Alberto et al. (2019). “RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli K-12*”. In: *Nucleic Acids Research* 47, pp. 212–220. DOI: 10.1093/nar/gky1077.
- Schmidt, Alexander et al. (Jan. 2016). “The quantitative and condition-dependent *Escherichia coli* proteome”. en. In: *Nature Biotechnology* 34.1, pp. 104–110. DOI: 10.1038/nbt.3418.
- Schneider, T D and R M Stephens (Oct. 1990). “Sequence logos: a new way to display consensus sequences.” In: *Nucleic Acids Research* 18.20, pp. 6097–6100.
- Schweizer, H, W Boos, and T J Larson (1985). “Repressor for the sn-glycerol-3-phosphate regulon of *Escherichia coli* K-12: cloning of the glpR gene and identification of its product.” en. In: *Journal of Bacteriology* 161.2, pp. 563–566. DOI: 10.1128/JB.161.2.563-566.1985.
- Seoh, H. K. and P. C. Tai (Mar. 1999). “Catabolic repression of secB expression is positively controlled by cyclic AMP (cAMP) receptor protein-cAMP complexes at the transcriptional level”. eng. In: *Journal of Bacteriology* 181.6, pp. 1892–1899.
- Sharon, E. et al. (2012). “Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters”. In: *Nature Biotechnology* 30.6, pp. 521–30. DOI: 10.1038/nbt.2205.

- Stormo and D. S. Fields (1998). “Specificity, free energy and information content in protein-DNA interactions”. In: *Trends Biochem Sci* 23.3, pp. 109–13. DOI: 10.1016/S0968-0004(98)01187-6.
- Stuart, Tim and Rahul Satija (2019). “Integrative single-cell analysis”. In: *Nature Reviews Genetics* 20, pp. 257–272. DOI: 10.1038/s41576-019-0093-7.
- Tareen, A. and Kinney (2019). “Biophysical models of cis-regulation as interpretable neural networks”. In: *bioRxiv*.
- Urtecho, G. et al. (2019). “Systematic Dissection of Sequence Elements Controlling sigma70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*”. In: *Biochemistry* 58.11, pp. 1539–1551. DOI: 10.1021/acs.biochem.7b01069.
- Urtecho, Guillaume et al. (2020). “Genome-wide Functional Characterization of *Escherichia coli* Promoters and Regulatory Elements Responsible for their Function.” In: *bioRxiv*. DOI: 10.1101/2020.01.04.894907. eprint: <https://www.biorxiv.org/content/early/2020/01/06/2020.01.04.894907.full.pdf>.
- Vilar, J. M. and Leibler (2003). “DNA looping and physical constraints on transcription regulation”. In: *J Mol Biol* 331.5, pp. 981–9. DOI: 10.1016/S0022-2836(03)00764-2.
- Vilar, J. M. and L. Saiz (2013). “Reliable prediction of complex phenotypes from a modular design in free energy space: an extensive exploration of the lac operon”. In: *ACS Synth Biol* 2.10, pp. 576–86. DOI: 10.1021/sb400013w.
- Yamamoto, Natsuko et al. (2009). “Update on the Keio collection of *Escherichia coli* single-gene deletion mutants”. In: *Molecular Systems Biology* 5, pp. 335–335. DOI: 10.1038/msb.2009.92.

Appendix A

EXTENDED EXPERIMENTAL DETAILS

Extended details of experimental design

Choosing target genes for Reg-Seq

Genes in this study were chosen to cover several different categories. 29 genes had some information on their regulation already known to validate our method under a number of conditions. 37 were chosen because the work of Schmidt, Kochanowski, Vedelaar, Ahrné, et al., 2016 demonstrated that gene expression changed significantly under different growth conditions. A handful of genes such as *minC*, *maoP*, or *fdhE* were chosen because we found either their physiological significance interesting, as in the case of the cell division gene *minC* or that we found the gene regulatory question interesting, such for the intra-operon regulation demonstrated by *fdhE*. The remainder of the genes were chosen because they had no regulatory information, often had minimal information about the function of the gene, and had an annotated transcription start site (TSS) in RegulonDB.

Choosing transcription start sites for Reg-Seq

A known limitation of the experiment is that the mutational window is limited to 160 bp. As such, it is important to correctly target the mutation window to the location around the most active TSS. To do this we first prioritized those TSS which have been extensively experimentally validated and catalogued in RegulonDB. Secondly we selected those sites which had evidence of active transcription from RACE experiments (Mendoza-Vargas et al., 2009) and were listed in RegulonDB. If the intergenic region was small enough, we covered the entire region with our mutation window. If none of these options were available, we used computationally predicted start sites.

Reg-Seq Sequencing

The total library was first sequenced by PCR amplifying the region containing the variant promoters as well as the corresponding barcodes. This allowed us to uniquely associate each random 20 bp barcode with a promoter variant. Any barcode which was associated with a promoter variant with insertions or deletions was removed from further analysis. Similarly, any barcode that was associated with multiple pro-

moter variants was also removed from the analysis. The paired end reads from this sequencing step were then assembled using the FLASH tool (Magoc and Salzberg, 2011). Any sequence with PHRED score less than 20 was removed using the FastX toolkit. Additionally, when sequencing the initial library, sequences which only appear in the dataset once were not included in further analysis in order to remove possible sequencing errors.

For all the MPRA experiments, only the region containing the random 20 bp barcode was sequenced, since the barcode can be matched to a specific promoter variant using the initial library sequencing run described above. For a given growth condition, each promoter yielded 20,000 to 500,000 usable sequencing reads. Under some growth conditions, genes were not analyzed further if they did not have at least 20,000 reads.

To determine which base pair regions were statistically significant a 99% confidence interval was constructed using the MCMC inference to determine the uncertainty.

Reg-Seq Growth conditions

The growth conditions studied in this study were inspired by (Schmidt, Kochanowski, Vedelaar, Ahrne, et al., 2015) and include differing carbon sources such as growth in M9 with 0.5% Glucose, M9 with acetate (0.5%), M9 with arabinose (0.5%), M9 with Xylose (0.5%) and arabinose (0.5%), M9 with succinate (0.5%), M9 with fumarate (0.5%), M9 with Trehalose (0.5%), and LB. In each case cell harvesting was done at an OD of 0.3. These growth conditions were chosen so as to span a wide range of growth rates, as well as to illuminate any carbon source specific regulators.

We also used several stress conditions such as heat shock, where cells were grown in M9 and were subjected to a heat shock of 42 degrees for 5 minutes before harvesting RNA. We grew in low oxygen conditions. Cells were grown in LB in a container with minimal oxygen, although some will be present as no anaerobic chamber was used. This level of oxygen stress was still sufficient to activate FNR binding, and so activated the anaerobic metabolism. We also grew cells in M9 with Glucose and 5mM sodium salicylate.

Growth with zinc was preformed at a concentration of 5mM ZnCl_2 and growth with

iron was preformed by first growing cells to an OD of 0.3 and then adding FeCl_2 to a concentration of 5mM and harvesting RNA after 10 minutes. Growth without cAMP was accomplished by the use of the JK10 strain which does not maintain its cAMP levels.

All knockout experiment were preformed in M9 with Glucose except for the knockouts for *arcA*, *hdfR*, and *phoP* which were grown in LB.

Reg-Seq Constructs

The following mutated constructs were integrated into the plasmid.

Genes	Forward Integration Site	Mutated Wild Type Sequence	5' RNA Sequence	Barcode
<i>fdoH,sdaB,thiM,yedI,ykgE,sdiA</i>	TTCGTCCTTCACCTCGAGCAGCTTATTCGTCGCTGTAT	TACCTTTGATTGCTGTGCCCTATTAGGCTTCTCCTCAGCGCTAGTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>yqhC,yicI,ybjT,mtgA,aphA,bdcR</i>	TTCGTCCTTCACCTCGAGCAGCTTTGCTTCAGTCAGATTGCG	GTTCAATCACTGAATCCGGTATTAGGCTTCTCCTCAGCGCTCCTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>yncD,rumB,yagH,eco,yfhG,htrB</i>	TTCGTCCTTCACCTCGAGCAGCTCGAGTCTATGTAACCGT	CAGGGGCTGTCATATCTCATATTAGGCTTCTCCTCAGCGGACTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>iap,yjiP,yedK,holC,aegA,rapA</i>	TTCGTCCTTCACCTCGAGCAGCTAAGATGGAAGCCGGGATA	CACCTCATAGAGCTGTGGAATATTAGGCTTCTCCTCAGCGTCGGTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>dusC,fdhE,dnaE,ycgB,yehS,yeiQ</i>	TTCGTCCTTCACCTCGAGCAGCTGTCGCAACATGATCTAC	CGGTTCTAGTCATGTTTGCTATTAGGCTTCTCCTCAGCGCAATCACTGGCCGTCGTTTACATGACTGACTGA		
<i>ydhO,hslU,ymgG,rlmA,modE,ycbZ</i>	TTCGTCCTTCACCTCGAGCAGCTGCTAAGTCACACTGTTGG	TTGTAATATCTGTCGCGGATTAGGCTTCTCCTCAGCGATCTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>yajL,yecE,ybiP,ybjL,ygdH,pcm</i>	TTCGTCCTTCACCTCGAGCAGCTCTAAACAGTTAGGCCGAGG	TTATGTTCACAACTGGCGTGTATTAGGCTTCTCCTCAGCGAGTTCAGTGGCCGTCGTTTACATGACTGACTGA		
<i>rcsf,sbcB,ygeR,mscK,mscS,ynal</i>	TTCGTCCTTCACCTCGAGCAGCTTTTATACTTGCTGCGG	TGGAAGTATTGGCTTTGTATTAGGCTTCTCCTCAGCGAGTATCACTGGCCGTCGTTTACATGACTGACTGA		
<i>ybdG,hicB,arcB,minC,ybeZ,ydjA</i>	TTCGTCCTTCACCTCGAGCAGCAGCGATCAATCAACTT	TATAGTTCTCCATGCACCTATTAGGCTTCTCCTCAGCGTGGTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>yggW,acul,yehU,yehT,ybiO,mscL</i>	TTCGTCCTTCACCTCGAGCAGCTTCGGATAGACTCAGGAAGC	ACAATAGACAGACCATGCATATTAGGCTTCTCCTCAGCGGCTTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>zapB,waaA-coaD,coaA,yjiJ,groSL,pyrLBI</i>	TTCGTCCTTCACCTCGAGCAGCTTATGATAGATTCGCTCGC	GAGTCGAGCTAGCATAGGAGTATTAGGCTTCTCCTCAGCGAATTCAGTGGCCGTCGTTTACATGACTGACTGA		
<i>RplKAIJ-rpoBC,yodB,atpI,msyB,ndk,thrLABC</i>	TTCGTCCTTCACCTCGAGCAGCTTTTCTACTTCCGGCTTGC	TTGTGGGAGCTCTTACCATATTAGGCTTCTCCTCAGCGAATTCAGTGGCCGTCGTTTACATGACTGACTGA		
<i>tig,tff,maoP,poxB,rsfA,mscM</i>	TTCGTCCTTCACCTCGAGCAGCTATTTGGGGTCTGAC	TCGTACGGGAATGACCATAGTATTAGGCTTCTCCTCAGCGTAATCACTGGCCGTCGTTTACATGACTGACTGA		
<i>arcA,tar,dpiBA,araAB,araC,xylF</i>	TTCGTCCTTCACCTCGAGCAGCTGACAATAGTTGAGCCCTT	AGACACAAGTAGCCGATTATTAGGCTTCTCCTCAGCGGTTTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>xylA,dicA,dicC,dicB,ompR,xapAB</i>	TTCGTCCTTCACCTCGAGCAGCTGTAATGTGTGT	CGGACTAAAGGATCAGTCATATTAGGCTTCTCCTCAGCGGCTTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>ilvC,asnA,jdnK,dinJ,yjiY,motAB</i>	TTCGTCCTTCACCTCGAGCAGCTATACGTAAGGGTCCGA	CATCGGATAACACAAGCGTTATTAGGCTTCTCCTCAGCGGCTTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>ftsK,cra,uwrD,adiY,znuCB,znuA</i>	TTCGTCCTTCACCTCGAGCAGCTTATGATGTCGGATACCCG	GATGTATACCTCCAGCTGGTTATTAGGCTTCTCCTCAGCGACCTCACTGGCCGTCGTTTACATGACTGACTGA		
<i>zupT,pitA,ecmB,leuABCD</i>	TTCGTCCTTCACCTCGAGCAGCTTGAATAACACGGGTCC	TGAGATATGTACTGGTGCCTATTAGGCTTCTCCTCAGCGATTGTCACTGGCCGTCGTTTACATGACTGACTGA		

Figure A.1: Promoter constructs for Reg-Seq. We show examples of the constructs generated for the Reg-Seq project. Here we display the different 5' mRNA sequences that follow the mutated region for each gene. The mutated region will be 160 base pairs, with 115 base pairs upstream and 45 base pairs downstream of the TSS for that gene listed in Table A.1.

Gene	Start Site	Transcription Direction
fdoH	4085867	rev
sdaB	2928035	fwd
thiM	2185451	rev
yedJ	2033449	rev
ykgE	321511	fwd
sdiA	1996867	rev
yqhC	3155262	rev
yicI	3836664	rev
ybjT	909320	rev
mtgA	3350504	rev
aphA	4269355	fwd
bdcR	4474096	fwd
yncD	1523276	rev
rumB	897947	fwd
yagH	285350	fwd
eco	2303851	fwd
yfhG	2690181	rev
htrB	1116709	rev
iap	2876547	fwd
ygjP	3235915	fwd
yedK	2009866	fwd
holC	4484273	rev
aegA	2585570	rev
rapA	63358	rev
dusC	2230395	rev
fdhE	4081359	rev
dnaE	197026	fwd
ycgB	1237285	rev
yehS	2212241	rev
yeiQ	2266214	rev
ydhO	1734357	fwd
hslU	4122354	rev
ymgG	1223097	rev
rlmA	1907086	rev
modE	794644	rev
ycbZ	1018330	rev
yajL	443748	rev
yecE	1950778	fwd
ybiP	904523	rev
ybjL	889945	rev
ygdH	2926272	fwd
pcm	2870686	rev
rcsF	220022	rev
sbcB	2082728	fwd
ygeR	2999918	rev
mscK	486492	fwd

Gene	Start Site	Transcription Direction
mscS	3069871	rev
ynaI	1395973	rev
ybdG	604684	rev
hicB	1509221	fwd
arcB	3353049	rev
minC	1226139	rev
ybeZ	693469	rev
ydjA	1848700	rev
yggW	3096620	fwd
acuI	3403446	fwd
yehU	2214673	rev
yehT	2212969	rev
ybiO	845736	rev
mscL	3438001	fwd
zapB	4118427	fwd
WaaA-coaD	3808516	fwd
coaA	4175107	rev
yjjJ	4621716	fwd
groSL	4370616	fwd
pyrLBI	4472553	rev
rplKAJL-rpoBC	4178354	fwd
yodB	2042294	fwd
atpIBEFHAGDC	3922525	rev
msyB	1114213	rev
ndk	2644913	rev
thrLABC	148	fwd
tig	455077	fwd
tff-rpsB-tsf	189712	fwd
maoP	3948058	fwd
poxB	911076	rev
rspA	1655186	rev
mscM	4390638	rev
arcA	4640508	rev
tar	1972716	rev
dpiBA	652172	fwd
araAB	70075	rev
araC	70241	fwd
xylF	3731069	fwd
xylA	3730807	rev
dicA	1647979	fwd
dicC	1647876	rev
dicB	1649597	fwd
ompR	3536707	rev
xapAB	2524910	rev
ilvC	3957912	fwd
asnA	3927129	fwd

Gene	Start Site	Transcription Direction
idnK	4494597	fwd
dinJ	246533	rev
yjiY	4591397	rev
motAB-cheAW	1977302	rev
ftsK	933138	fwd
cra	87969	fwd
uvrD	3997907	fwd
adiY	4338042	rev
znuCB	1942634	fwd
znuA	1942661	rev
zupT	3182433	fwd
pitA	3637612	fwd
ecnB	4376509	fwd
leuABCD	83735	rev

Table A.1: All TSS for all genes investigated in Reg-Seq.

There are 160 base pairs of mutated sequence for each regulatory region, 115 base pairs upstream and 45 base pairs down stream of the transcription start site. The TSS location is shown, as well as whether the gene is transcribed in a 5' to 3' direction or a 3' to 5' direction.

EXTENDED ANALYSIS DETAILS

B.1 Validating Reg-Seq against previous methods and results

Reg-Seq is effectively a third-generation of the use of Sort-Seq methods for the discovery of regulatory architecture. The primary difference between the present work and previous generations (Kinney, Murugan, et al., 2010; Belliveau et al., 2018) is the use of RNA-Seq rather than fluorescence and cell sorting as a readout of the level of expression of our promoter libraries. As such, there are many important questions to be asked about the comparison between the earlier methods and this work. We attack that question in several ways. First, as shown in Figure B.1, we have performed a head-to-head comparison of the two approaches to be described further in this section. Second, as shown in the next section, our list of candidate promoters included roughly 20% for which the community has some knowledge of their regulatory architecture. In these cases, we examined the extent to which our methods recover the known features of regulatory control about those promoters.

Comparison between Reg-Seq by RNA-Seq and fluorescent sorting

As the basis for comparing the results of the fluorescence-based Sort-Seq approach with our RNA-Seq-based approach, we use information footprints, expression shifts and sequence logos as our metrics. Figure B.1 shows examples of this comparison for four distinct genes of interest. Figure B.1(A) shows the results of the two methods for the *lacZYA* promoter with special reference to the CRP binding site. Both the information footprint and the sequence logo identify the same binding site.

Figure B.1(B) provides a similar analysis for the *dgoRKADT* promoter where once again the information footprints and the sequence logos from the two methods are in reasonable accord. Figure B.1(C) provides a quantitative dissection of the *relBE* promoter which is repressed by RelBE. Here we use both information footprints and expression shifts as a way to quantify the significance of mutations to different binding sites across the promoter. Finally, Figure B.1(D) shows a comparison of the two methods for the *marRAB* promoter. The two approaches both identify a MarR binding site.

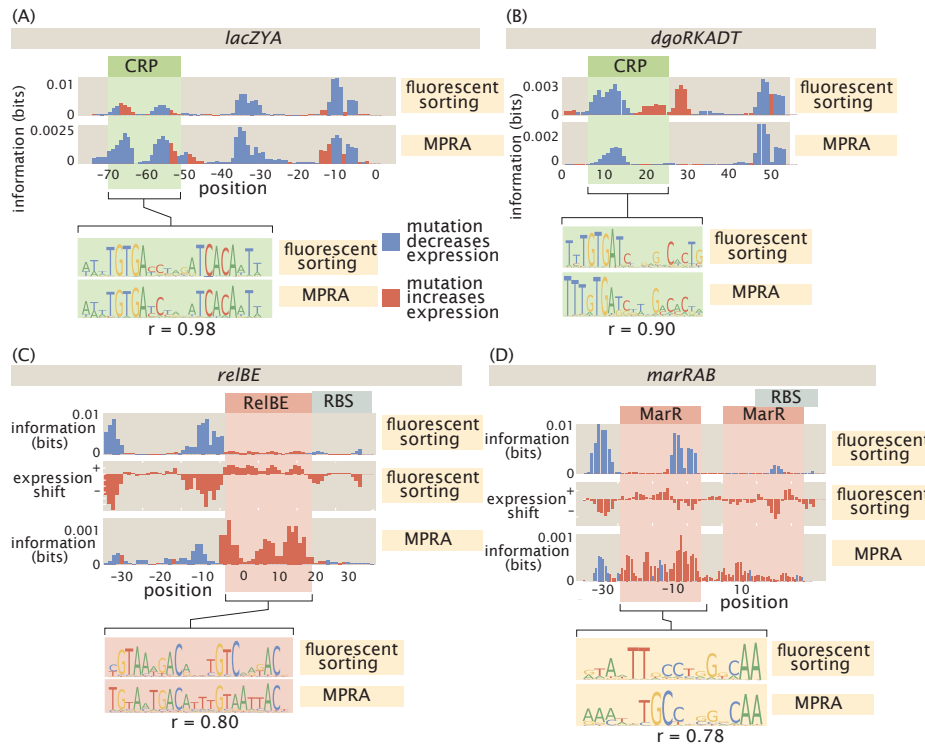


Figure B.1: A summary of four direct comparisons of measurements using fluorescence and sorting and using RNA-Seq. (A) CRP binds upstream of RNAP in the *lacZYA* promoter. Despite the different measurement techniques for the two inferred energy matrices and their corresponding sequence logos, the CRP binding sites have a Pearson correlation coefficient of $r = 0.98$. (B) The *dgoRKADT* promoter is activated by CRP in the presence of galactonate. The FACS measurements were taken in the JK10 strain in the presence of 500mM cAMP. In both cases, a type II activator binding site can be identified based on the signals in the information footprint in the area indicated in green. Additionally the quantitative agreement between the CRP binding preference matrices are strong, with $r = 0.9$. (C) The *relBE* promoter is repressed by RelBE. The inferred matrices between the two measurement methods have $r = 0.8$. (D) The *marRAB* promoter is repressed by MarR. The features we can observe in the information footprint reflect this under measurement with both FACS or RNA-seq. The inferred energy matrices (data not shown) and sequence logos shown have $r = 0.78$. The right most MarR site overlaps with a ribosome binding site. The overlap has a stronger obscuring effect on the sequence specificity of the FACS measurement, which measures protein levels directly, than it does on the output of the RNA-seq measurement.

Ability of Reg-Seq to recover known regulatory architectures

In total, we have tested over 20 genes for which there is already some substantial regulatory knowledge reported in the literature. The successes and failures of this test are detailed in Figure B.2. For those promoters which have strong evidence of a binding site, as determined by RegulonDB (Santos-Zavaleta et al., 2019), we recover all relevant transcription factor binding sites for 12 out of 16 cases, the majority of relevant binding sites for 2 out of 16 cases, and miss all or most of the regulation for just 2 promoters. We identify a total of 22 previously known high evidence binding sites.

These results showcase that our method largely agrees with the established literature but also highlights several areas in which our method is prone to missing regulatory elements. One failure mode is caused by the presence of strong secondary binding sites. For example, in the *araC* promoter, as shown in Figure B.2(C), the only binding signatures that appear in the information footprint are from a secondary RNAP site. The secondary site seems to be expressed constitutively, and in the cases where the primary start site is even partially repressed, the secondary start site will dominate transcription and obscure the many binding sites that are in this promoter.

If there are large numbers of regulatory elements, the data will often only show the few most important elements. If we look at the *marR* promoter in Figure B.2(C), we can only see the signature of the two MarR sites even though CpxR, Fis, and CRP are all known to bind to the promoter. MarR is a strong enough repressor that mutating any of the other transcription factor sites is unlikely to meaningfully change gene expression unless the MarR site is also mutated. This illustrates that the regulatory architectures discovered in this study represent a lower bound on what exists in each promoter.

Finally, for some genes such as *dicA* there was no known TSS prior to the experiment. Although there is a small regulatory region between *dicA* and its neighboring gene, this does not ensure that we will include the strongest RNAP sites. Better mapping of transcription start sites could improve our method.

We next consider low evidence binding sites. Other research determined the loca-

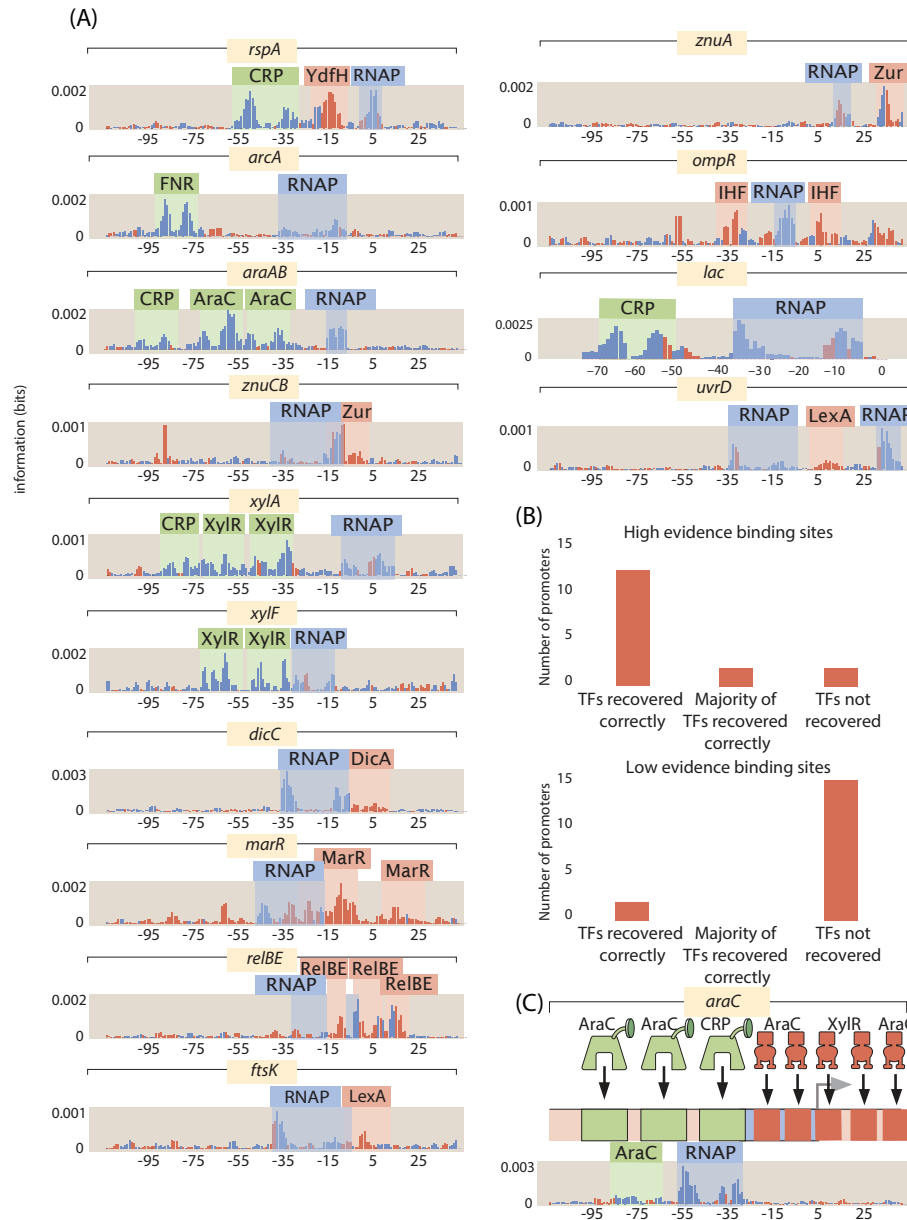


Figure B.2: Reg-Seq analysis of “gold standard” promoters. (A) Information footprints for known and properly recovered binding sites. (B) A summary of how well the Reg-Seq results conform to literature results. The sites that are low evidence in the literature are determined by RegulonDB (Santos-Zavaleta et al., 2019). (C) The information footprint and known binding sites for the *araC* promoter. Despite all the binding sites present, the only binding signature that appears is for RNAP.

tions of the low evidence sites through gene expression analysis and sequence comparison to consensus sequences (Compan and Touati, 1994; Kumar and Shimizu, 2011; Easton and Kushner, 1983). For 5 promoters in our list, the binding sites

location itself is not known, only that the TF in question regulates the gene. For these promoters we recover the known regulation in only 2 out of 15 cases. Comparison to consensus sequences can be unreliable and generate false positives when the entirety of the *E. coli* genome is considered. Gene expression analysis alone has difficulty ruling out indirect effects of a given transcription factor on gene expression and regulation determined by this method may occur outside of the 160 bp mutation window we consider. As our results recover high evidence sites well, the poor recovery of sites based on sequence gazing and gene expression analysis most likely indicates that these methods are unreliable for determining binding locations.

We note that the first aim of our methods is regulatory discovery. We would like to be able to determine how previously uncharacterized promoters are regulated and ultimately, this is a question of binding-site and transcription factor identification. For that task, we do not require perfect correspondence between the two methods. With regulatory sites identified, our next objective is the determination of energy matrices that will allow us to turn binding site strength into a tunable knob that can nearly continuously tune the strength of transcription factor binding, thus altering gene expression in predictable ways as already shown in our earlier work in Chapter 2 (Barnes et al., 2019). The r -values between energy matrices range from 0.78 to 0.96, indicating reasonable to very good agreement. Reg-Seq appears to be, if anything, more accurate than previous methods as it has higher relative information content in known areas of transcription factor binding and also does not have repressor-like bases on CRP sites as in Figure B.1(A) and (B).

B.2 Information footprints

We use information footprints as a tool for hypothesis generation to identify regions which may contain transcription factor binding sites. In general, a mutation within a transcription factor site is likely to severely weaken that site. We look for groups of positions where mutation away from wild type has a large effect on gene expression. Our data sets consist of nucleotide sequences, the number of times we sequenced the construct in the plasmid library, and the number of times we sequenced its corresponding mRNA. A simplified data set on a 4 nucleotide sequence then might look like

Sequence	Library Sequencing Counts	mRNA Counts
ACTA	5	23
ATTA	5	3
CCTG	11	11
TAGA	12	3
GTGC	2	0
CACA	8	7
AGGC	7	3

$$I_b = \sum_{m=0}^1 \sum_{\mu=0}^1 p(m, \mu) \log_2 \left(\frac{p(m, \mu)}{p_{mut}(m) p_{expr}(\mu)} \right). \quad (\text{B.1})$$

$p_{mut}(m)$ in equation B.1 refers to the probability that a given sequencing read will be from a mutated base. $p_{expr}(\mu)$ is a normalizing factor that gives the ratio of the number of DNA or mRNA sequencing counts to total number of counts.

The mutual information quantifies how much a piece of knowledge reduces the entropy of a distribution. At a position where base identity matters little for expression level, there would be little difference in the frequency distributions for the library and mRNA transcripts. The entropy of the distribution would decrease only by a small amount when considering the two types of sequencing reads separately.

We are interested in quantifying the degree to which mutation away from a wild type sequence affects expression. Although there are obviously 4 possible nucleotides, we can classify each base as either wild-type or mutated so that b in equation B.1 represents only these two possibilities.

If mutations at each position are not fully independent, then the information value calculated in equation B.1 will also encode the effect of mutation at correlated positions. If having a mutation at position 1 is highly favorable for gene expression and is also correlated with having a mutation at position 2, mutations at position 2 will also be enriched amongst the mRNA transcripts. Position 2 will appear to have high mutual information even if it has minimal effect on gene expression. Due to the DNA synthesis process used in library construction, mutation in one position can

make mutation at other positions more likely by up to 10 percent. This is enough to cloud the signature of most transcription factors in an information footprint calculated using equation B.1.

We need to determine values for $p_i(m|exp)$ when mutations are independent, and to do this we need to fit these quantities from our data. We assert that

$$\langle mRNA \rangle \propto e^{-\beta E_{eff}} \quad (B.2)$$

is a reasonable approximation to make. $\langle mRNA \rangle$ is the average number of mRNAs produced by that sequence for every cell containing the construct and E_{eff} is an effective energy for the sequence that can be determined by summing contributions from each position in the sequence. There are many possible underlying regulatory architectures, but to demonstrate that our approach is reasonable let us first consider the simple case where there is only a RNAP site in the studied region. We can write down an expression for average gene expression per cell as

$$\langle mRNA \rangle \propto p_{bound} \propto \frac{\frac{p}{N_{NS}} e^{-\beta E_P}}{1 + \frac{p}{N_{NS}} e^{-\beta E_P}} \quad (B.3)$$

where p_{bound} is the probability that the RNAP is bound to DNA and is known to be proportional to gene expression in *E. coli* (Garcia and Phillips, 2011), E_P is the energy of RNAP binding, N_{NS} is the number of nonspecific DNA binding sites, and p is the number of RNAP. If RNAP binds weakly then $\frac{p}{N_{NS}} e^{-\beta E_P} \ll 1$. We can simplify equation B.3 to

$$\langle mRNA \rangle \propto e^{-\beta E_P}. \quad (B.4)$$

If we assume that the energy of RNAP binding will be a sum of contributions from each of the positions within its binding site then we can calculate the difference in gene expression between having a mutated base at position i and having a wild type base as

$$\frac{\langle mRNA_{WT_i} \rangle}{\langle mRNA_{Mut_i} \rangle} = \frac{e^{-\beta E_{P_{WT_i}}}}{e^{-\beta E_{P_{Mut_i}}}} \quad (B.5)$$

$$\frac{\langle mRNA_{WT_i} \rangle}{\langle mRNA_{Mut_i} \rangle} = e^{-\beta(E_{P_{WT_i}} - E_{P_{Mut_i}})}. \quad (B.6)$$

In this example we are only considering single mutation in the sequence so we can further simplify the equation to

$$\frac{\langle mRNA_{WT_i} \rangle}{\langle mRNA_{Mut_i} \rangle} = e^{-\beta \Delta E_{P_i}}. \quad (B.7)$$

We can now calculate the base probabilities in the expressed sequences. If the probability of finding a wild type base at position i in the DNA library is $p_i(m = WT|exp = 0)$ then

$$p_i(m = WT|exp = 1) = \frac{p_i(m = WT|exp = 0) \frac{\langle mRNA_{WT_i} \rangle}{\langle mRNA_{Mut_i} \rangle}}{p_i(m = Mut|exp = 0) + p_i(m = WT|exp = 0) \frac{\langle mRNA_{WT_i} \rangle}{\langle mRNA_{Mut_i} \rangle}} \quad (B.8)$$

$$p_i(m = WT|exp = 1) = \frac{p_i(m = WT|exp = 0) e^{-\beta \Delta E_{P_i}}}{p_i(m = Mut|exp = 0) + p_i(m = WT|exp = 0) e^{-\beta \Delta E_{P_i}}}. \quad (B.9)$$

Under certain conditions, we can also infer a value for $p_i(m|exp = 1)$ using a linear model when there are any number of activator or repressor binding sites. We will demonstrate this in the case of a single activator and a single repressor, although a similar analysis can be done when there are greater numbers of transcription factors. We will define $P = \frac{p}{N_{NS}} e^{-\beta E_P}$. We will also define $A = \frac{a}{N_{NS}} e^{-\beta E_A}$ where a is the number of activators, and E_A is the binding energy of the activator. We will finally define $R = \frac{r}{N_{NS}} e^{-\beta E_R}$ where r is the number of repressors and E_R is the binding energy of the repressor. We can write

$$\langle mRNA \rangle \propto p_{bound} \propto \frac{P + PAe^{-\beta \epsilon_{AP}}}{1 + A + P + R + PAe^{-\beta \epsilon_{AP}}}. \quad (B.10)$$

If activators and RNAP bind weakly but interact strongly, and repressors bind very strongly, then we can simplify equation B.10. In this case $A \ll 1$, $P \ll 1$, $PAe^{-\epsilon_{AP}} \gg P$, and $R \gg 1$. We can then rewrite equation B.10 as

$$\langle mRNA \rangle \propto \frac{PAe^{-\beta\epsilon_{AP}}}{R} \quad (\text{B.11})$$

$$\langle mRNA \rangle \propto e^{-\beta(-E_P - E_A + E_R)}. \quad (\text{B.12})$$

As we typically assume that RNAP binding energy, activator binding energy, and repressor binding can all be represented as sums of contributions from their constituent bases, the combination of the energies can be written as a total effective energy E_{eff} which is a sum of contributions from all positions within the binding sites.

We fit the parameters for each base using a Markov Chain Monte Carlo Method. Two MCMC runs are conducted using randomly generated initial conditions. We require both chains to reach the same distribution to prove the convergence of the chains. We do not wish for mutation rate to affect the information values so we set the $p(WT) = p(Mut) = 0.5$ in the information calculation. The information values are smoothed by averaging with neighboring values.

B.3 Estimating mutual information from observed data and model predictions

One difficulty with estimating the mutual information from model predictions is that base pair identity A, C, G, T and the gene expression level μ are both discrete variables, while binding energy predictions from the model (x) is a continuous variable. Formally, the mutual information is given by

$$I(\mu, x) = \int_{x=-\infty}^{x=+\infty} dx \sum_{\mu} p(x, \mu) \log_2 \left(\frac{p(x, \mu)}{p(x)p(\mu)} \right), \quad (\text{B.13})$$

where μ is a measure of the gene expression and is equal to the sorting bin number 1, 2, 3, 4 in the case of Sort-Seq and

$$\mu = \begin{cases} 0, & \text{for sequencing reads from the DNA library} \\ 1, & \text{for sequencing reads originating from mRNA,} \end{cases} \quad (\text{B.14})$$

during Reg-Seq (discussed in Chapter 4).

The probability distribution p is not one that we have full access to, as we only have a discrete set of predictions (one from each of the N unique DNA sequences in our data set). To compensate for the issues that can arise from estimating a continuous distribution from discrete data, we make use of the fact that any transformation that preserves the rank order z_q (for instance multiplying all model predictions by a constant) the mutual information is unchanged.

We will define z_q as the rank order in binding energy predictions of the q th sequence. We estimate $I(\mu, z)$ by first calculating binding energy predictions

$$x = \sum_{i=1}^L \sum_{j=A}^T \theta_{ij} \cdot s_{ij}, \quad (\text{B.15})$$

and then converting them to a rank order predictions z .

We then discretize the energy predictions into 1000 "bins" and convolve with a Gaussian kernel to estimate the probability distribution $p(z, \mu)$. We can then calculate the mutual information with

$$I(\mu, z)_{\text{smoothed}} = \sum_{z=1}^{1000} \sum \mu F(\mu, z) \log_2 \frac{F(\mu, z)}{F(z) \cdot F(\mu)}, \quad (\text{B.16})$$

where $F(\mu, z)$ is the probability distribution $p(\mu, z)$ estimated from finite data.

B.4 Markov Chain Monte Carlo fitting procedure

We will often want to infer parameter values under conditions where posterior distributions are difficult to work with. In these difficult cases, Markov Chain Monte Carlo (MCMC) methods can be used. MCMC gets its name from two processes, Monte Carlo and Markov Chain. Monte Carlo is a method for estimating features of a distribution by randomly drawing samples from the distribution. For example, one could estimate the mean or standard

deviation of a distribution by drawing random samples and computing the mean and standard deviation for those samples. Fig. B.3 (A) schematizes this process. Markov Chain is a type of algorithm for drawing random samples. In this Monte Carlo method, schematized in Fig. B.3 (B), a random value is added to current parameter values to generate the next sample. In a Markov Chain there is no

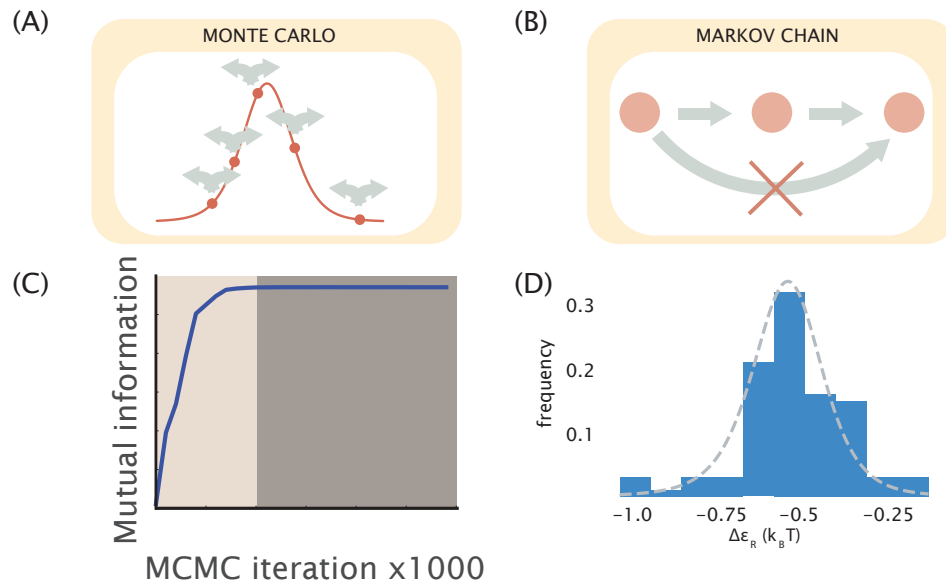


Figure B.3: Parameter inference using Markov Chain Monte Carlo. (A) A “Monte Carlo” technique refers to the process of taking random samples from a distribution and using these samples to estimate properties of the distribution. (B) A “Markov Chain” refers to a process of generating random samples in which an operation is performed only on the current sample to generate the next sample. The identity of each sample depends only on the sample immediately preceding it, and lacks memory of previous samples. (C) A short MCMC example chain for inferring a parameter. (D) A histogram generated from the chain plotted in (C).

memory of any previous samples; new steps are calculated only with information from the previous step. Each new step is accepted or rejected based on the relative likelihood of the new model compared to the old model. Successive iterations of MCMC will increase the mutual information to a plateau as shown in B.3 (C). This early period is known as a “burn in” period and is discarded from the analysis. Further samples can be used to determine the posterior distribution of the parameter, an example of which is shown in B.3 (D).

A full discussion of Markov Chain Monte Carlo method is beyond the scope of this study, but here we will provide a brief explanation. A full explanation of the method can be found in Neal, 1993. Markov Chain Monte Carlo (MCMC) methods are often used when functions are not amenable to analytical solutions or calculations. MCMC methods allow the expectation value of a given parameter, and its uncertainty without requiring us to have full access to the underlying probability distribution. As with many cases in biology, the true underlying probability distribution is often complicated and difficult to access.

As proven by Kinney, 2008, the likelihood of a model that predicts gene output is

$$L(\theta|\mu_s) \propto 2^{NI_{smooth}(\mu,z)}, \quad (\text{B.17})$$

where N is the total number of independent sequences, I_{smooth} is the smoothed mutual information between gene expression (μ) and DNA sequence (z).

The probability distributions $p(\theta)$ is very difficult to handle analytically. The reason why we use MCMC is that you can estimate properties using the target probability distribution without needing to know the distribution. For example, we can estimate $\langle\theta\rangle$ by drawing many samples of θ using MCMC and taking the mean of the parameters.

We therefore need to construct a Markov Chain whose stationary distribution converges to the distribution of interest $p(\theta)$. A Markov chain is a sequence of values $\theta_1, \theta_2, \theta_3, \dots, \theta_N$ for N steps. We can then find $\langle\theta\rangle$ with

$$\langle\theta\rangle = \frac{\sum_{N=1}^{100} \theta_N}{N}, \quad (\text{B.18})$$

A Markov chain has no memory. That is the probability that the N th value in the chain takes a value θ_N depends only on the $(N - 1)$ th value in the chain. To make things a bit more concrete, let's leave aside θ for the time being. Imagine that we have a light switch, and we know the switch is "on" 25% of the time and "off" 75% of the time. For each "step" in our Markov chain, we can change the state of the switch; if the switch is on, we turn it off with some rate k_{off} , and if the switch is currently off, we turn it on with the rate k_{on} . A sequence of states will be generated. One example would be off, off, off, on, on, on, off. These states constitute a Markov chain and if the chain is continued for long enough, the stationary distribution will converge such that $p_{on} = 0.25$ and $p_{off} = 0.75$ as we knew going in.

A Markov chain is stationary if detailed balance is satisfied between its states. The condition of detailed balance obtains if the total rate of transitions from on to off is the same as the total rate of transitions from off to on. Mathematically, this condition can be written as

$$k_{on} \times p_{off} = k_{off} \times p_{on}. \quad (\text{B.19})$$

The above equation allows calculation of k_{on} and k_{off} .

$$\frac{k_{on}}{k_{off}} = \frac{p_{on}}{p_{off}} = \frac{1}{3}. \quad (\text{B.20})$$

As long as the ratio of transition rates is satisfied and enough steps are taken, then the stationary distribution will converge to the proper distribution.

For the far more complicated task of estimating $p(\theta)$ we can fall back on the standard Metropolis-Hastings sampling algorithm. θ will be a matrix where θ_{ij} will be the energetic contributions to the energy matrix for the i th position and the j th base pair where $i \in 1, 2, 3, \dots, L$ and $j \in A, C, G, T$. We can then follow the procedure.

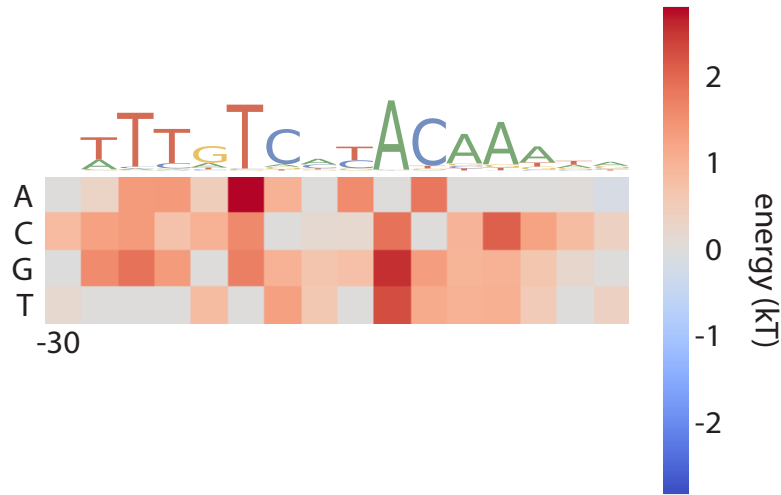


Figure B.4: An example energy matrix for the YieP binding site of the *ykgE*. An example energy matrix that describes the binding energy between the transcription factor YieP and DNA is shown. By convention, the wild-type nucleotides have zero energy, and each other entry represents the change in binding energy upon a mutation from the wild type nucleotide to a new nucleotide at that position.

1. Start with a random energy matrix θ_0 .
2. Make a random perturbation $d\theta$ to θ_0 . This perturbation will have a small adjustment to each element of θ .
3. Compute the model likelihoods $L(\theta_0)$ and $L(\theta_0 + d\theta)$ using equation B.17.
4. If $L(\theta_0 + d\theta) > L(\theta)$, accept the new parameter values $\theta_0 + d\theta$ as the next element in the markov chain (θ_1). Otherwise accept $\theta_0 + d\theta$ with probability $\frac{L(\theta_0 + d\theta)}{L(\theta_0)}$. If the step is rejected, the next element θ_1 in the Markov chain reset

to its previous value (θ_0). The acceptance/rejection probabilities mean that detailed balance is satisfied between the states θ_0 and $\theta_0 + d\theta$.

5. Repeat steps 2-4 until the chain converges to the stationary distribution. In practice this can be determined by monitoring when the mutual information plateaus, as can be seen in Fig. B.3 (C).
6. To be certain that the distribution has in fact converged to the proper stationary distribution, multiple Markov Chains should be run starting with step 1. If all the chains converge to the same distribution, then they have properly converged.

The end result of these model-fitting efforts is an optimized linear binding energy matrix like the one shown in Fig. B.4. You can get a measure of the uncertainty in θ_{ij} by forming a confidence interval out of the distribution of parameters formed from the Markov Chain. This inference is performed using the MPAtch software (Ireland and Kinney, 2016).

Uncertainty in Reg-Seq due to number of independent sequences

1400 promoter variants were ordered from TWIST Bioscience for each promoter studied. Due to errors in synthesis, additional mutations are introduced into the ordered oligos. As a result, the final number of variants received was an average of 2200 per promoter. To test whether the number of promoter variants is a significant source of uncertainty in the experiment we computationally reduced the number of promoter variants used in the analysis of the *zapAB* -10 RNAP region. Each sub-sampling was performed 3 times. The results, as displayed in Figure B.5, show that there is only a small effect on the resulting sequence logo until the library has been reduced to approximately 500 promoter variants.

B.5 TOMTOM motif comparison

In some cases, we used an alternative approach to mass spectrometry to discover the TF identity regulating a given promoter based on sequence analysis using a motif comparison tool. TOMTOM (Gupta et al., 2007) is a tool that uses a statistical method to infer if a putative motif resembles any previously discovered motif in a database. Of interest, it accounts for all possible offsets between the motifs. Moreover, it uses a suite of metrics to compare between motifs such as Kullback-Leibler

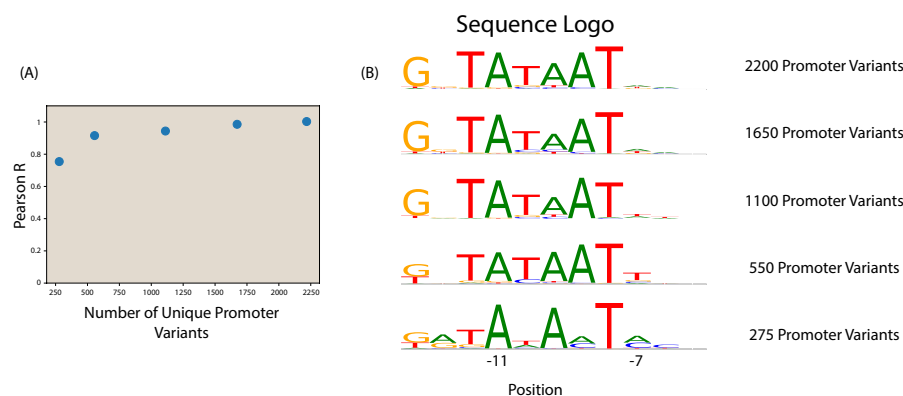


Figure B.5: A comparison of RNAP -10 site sequence logos. (A) This figure shows the Pearson correlation coefficient between the energy matrix models inferred from the full dataset (2200 unique promoter variants) and that from a computationally restricted dataset. (B) Sequence logos of the RNAP -10 region from each sub-sampled dataset.

divergence, Pearson correlation, and euclidean distance, among others.

We performed comparisons of the motifs generated from our energy matrices to those generated from all known transcription factor binding sites in RegulonDB. Figure B.6 shows a result of TOMTOM, where we compared the motif derived from the -35 region of the *ybjX* promoter and found a good match with the motif of PhoP from RegulonDB.

The information derived from this approach was then used to guide some of the TF knockout experiments, in order to validate its interaction with a target promoter characterized by the loss of the information footprint. Furthermore, we also used TOMTOM to search for similarities between our own database of motifs, in order to generate regulatory hypotheses in tandem. This was particularly useful when looking at the group of GlpR binding sites found in this experiment.

B.6 BioInformatic methods - TOMTOM motif comparison

There have been many attempts to scan the *E. coli* genome using consensus matrices built from the known binding sites of a transcription factor (Suzuki, 2003). However, as we see in section B.1, these efforts often fall short. One important reason is that a typical interaction energy between an activator and RNAP is $\approx -4 k_b T$ (Forcier et al., 2018). For CRP, the total protein DNA interaction energy is of a similar

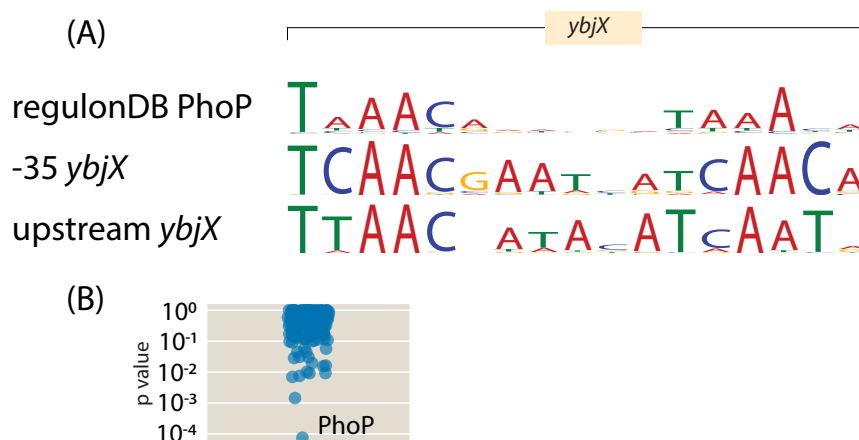


Figure B.6: Motif comparison using TOMTOM. Searching our energy motifs against the RegulonDB database using TOMTOM allowed us to guide our TF knockout experiments. Here we show the sequence logos of the PhoP transcription factor from RegulonDB (top) and the one generated from the ybjX promoter energy matrix. E-value = 0.01 using Euclidean distance as a similarity matrix.

magnitude. Without interacting with RNAP, a typical CRP site will have a difficult time being differentiated from the non specific background, as a binding site like the O3 operator for LacI only has a total binding energy of $\approx -10 k_b T$. O3 is typical of a weak specific binding site.

For future work, we are interested in using computational tools like TOMTOM to identify transcription factors. While, as we say, whole genome searches often cause a lot of false positive, we may get better results in the future by restricting analysis to only sites that we know are active transcription factor binding sites. Some examples of this type of search can be seen in Fig. 4.10 for FNR and GlpR binding sites found in the Reg-Seq project. We can draw a rough cutoff of a p-value of less than 10^{-4} . This cutoff does identify FNR and GlpR for these cases, and especially for the cases of FNR, the binding sites we found in the Reg-Seq project are very similar according to TOMTOM. This does imply that future attempts to "cluster" binding sites could be a useful tool, but they are not perfect. For example, CRP cannot be differentiated in any way from FNR. Not all binding sites cluster together well. The GlpR sites in particular only worked for a subset of the 5 sites. In future work, refinements of computational searches, possibly using deep learning methods, could be extremely useful for identification.

B.7 Sigma Factors

As discussed in the Introduction, each RNAP core protein must act in concert with a bacterial σ factor. The available σ factors are σ 70, σ 54, σ 38, σ 32, σ 28, σ E, and σ FecI. σ 70 is "housekeeping" σ factor and by far the most common σ factor. σ 38 is a general stress response σ factor in *E. coli*. Its copy number increases under stress and the onset of stationary phase, but it is available at lower levels under ordinary growth conditions. σ E is the envelope stress σ factor, and most relevantly, responds to heat shock conditions transiently.

σ 32 also responds to heat shock, while σ 54 responds to nitrogen starvation. σ 28 is a σ factor involved in flagellar synthesis. It competes with σ 70 and preferentially binds the RNAP core enzyme, but under ordinary growth conditions is only approximately half as prevalent as σ 70.

σ factor prevalence is a form of gene regulation in bacteria, generally at a wide scale. Even for a minor σ factor such as σ 32, there are approximately 100 known RNAP binding sites (Keseler et al., 2013). This is a similar scale to the number of known CRP or FNR binding sites (Keseler et al., 2013).

The heat shock σ factors have a large fold change when the cells are exposed to a five minute heat shock at 42°C. Additionally, they have relatively low leakiness. In contrast, other σ factors, such as σ 38, while also transiently upregulated under stress conditions, has a very high leakiness and is still quite active under exponential growth.

Despite using the same core enzyme when binding DNA as shown in B.7, the sequence binding preference can be quite dissimilar between different σ factors, allowing RNAP to discriminate between different binding sites. The consensus binding sequences are given in B.1.

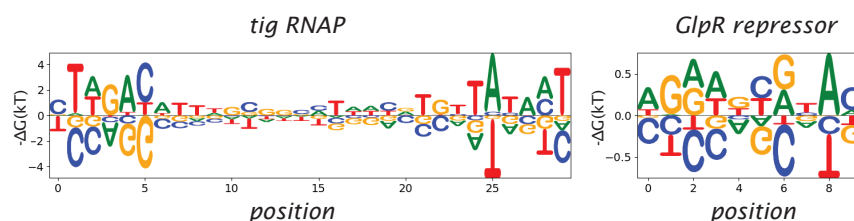


Figure B.7: Sequence logos for models fit using neural network fitting methods. These types of fitting procedures are far more computationally efficient, and will be important for the next generation of Reg-Seq.

σ Factor	Consensus Sequence
σ 70	TTGAC - \approx 15 bp spacer - TGNTATAAT
σ 38	TTGAC - \approx 14 bp spacer - TGTGCTATACT
σ 32	CTTGAA - \approx 15 bp spacer - CCCCATATAT
σ E	CCCCATtTa
σ H	CTGGCACA - \approx 3 bp spacer - ATTTGC(A/T)T
σ 28	GCCGATAA

Table B.1: All consensus σ factor binding sites

Consensus sites are taken from EcoCyc (Keseler et al., 2013)

We compare the consensus sequences for each σ factor to the binding energy matrices for each of the regulatory elements discovered. The most prevalent RNAP binding sites are found by human inspection and are

B.8 Binding sites regulating divergent operons

In addition to discovering new binding sites, we have discovered additional functions of known binding sites. In particular, in the case of *bdcR*, the repressor for the divergently transcribed gene *bdcA* (Partridge et al., 2009), is also shown to repress *bdcR* in Figure B.8(A). Similarly in Figure B.8(B) *IlvY* is shown to repress *ilvC* in the absence of inducer. Divergently transcribed operons that share regulatory regions are plentiful in *E. coli*, and although there are already many known examples of transcription factor binding sites regulating several different operons, there are almost certainly many examples of this type of transcription that have yet to be discovered.

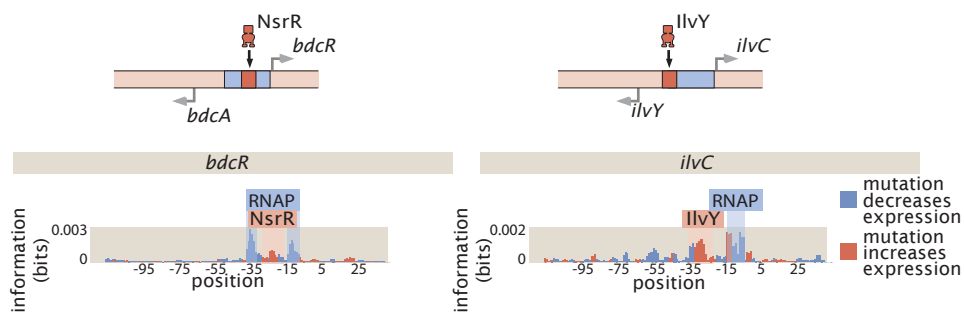


Figure B.8: Multipurpose binding sites. Two cases in which we see transcription factor binding sites that we have found to regulate both of the two divergently transcribed genes.

Multi-purpose binding sites allow for more genes to be regulated with fewer binding

Gene	Location compared to TSS	σ Factor
acuI	-105	σ 70
acuI	-25	σ 70
aegA	-9	σ 70
adiY	19	σ 70
aphA	-11	σ 70
araAB	-13	σ 70
araC	-32	σ 70
arcA	-99	σ 38
arcA	-12	σ 70
arcB	-11	σ 70
asnA	-10	σ 70
bdcR	-10	σ 70
coaA	-12	σ 70
cra	-12	σ 70
dicC	-11	σ 70
dinJ	-11	σ 70
dnaE	-11	σ 24
dpiBA	-24	σ 70
dusC	-23	σ 70
ecnB	-11	σ 70
fdhE	-12	σ 70
ftsK	-11	σ 38
groSL	-15	σ 32
groSL	19	σ 70
hicB	-11	σ 70
holC	-14	σ 32
hslU	-10	σ 32
htrB	-10	σ 38
iap	-12	σ 38
iap	33	σ 38
ilvC	-8	σ 70
maoP	-21	σ 70
minC	-12	σ 70
modE	-26	σ 70
mscK	-14	σ 38
mscL	-13	σ 38
mscM	35	σ 54
ompR	-11	σ 70
pcm	-11	σ 70
pit	-13	σ 70
poxB	-12	σ 38
rapA	-10	σ 70
rapA	-103	σ 38
rcsF	-62	σ 38
rcsF	-11	σ 70

Gene	Location compared to TSS	σ Factor
rlmA	-9	σ 70
rlmA	-64	σ 70
rspA	5	σ 38
rumB	-11	σ 70
sbcB	-11	σ 70
sdaB	-19	σ 70
sdiA	-54	σ 70
tff-rpsB-tsf	-12	σ 70
tff-rpsB-tsf	-84	σ 70
thiM	-11	σ 70
thrLABC	-14	σ 70
tig	-11	σ 70
uvrD	-12	σ 70
WaaA-coaD	-13	σ 70
xylA	-8	σ 70
xylF	-12	σ 70
ybdG	-11	σ 70
ybeZ	29	σ 70
ybeZ	-14	σ 32
ybiO	19	σ 38
ybiO	2	σ 70
ybjL	-11	σ 38
ybjL	-58	σ 70
ybjL	20	σ 70
ybjT	-12	σ 70
ybjT	10	σ 70
ycbZ	-8	σ 70
ycbZ	-11	σ 70
ycbZ	-25	σ 70
ycgB	-11	σ 38
ydhO	12	σ 70
ydjA	-13	σ 70
ydjA	17	σ 24
yecE	-1	σ 70
yecE	-33	σ 70
yedJ	-30	σ 70
yedJ	-11	σ 70
yedK	13	σ 70
yedK	25	σ 38
yehS	8	σ 70
yehT	-11	σ 70
yehT	12	σ 38

Gene	Location compared to TSS	σ Factor
yehU	-8	σ 70
yehU	36	σ 70
yehQ	-12	σ 70
yfhG	32	σ 70
ygdH	-12	σ 70
ygeR	-10	σ 70
yggW	-14	σ 32
yggP	-23	σ 70
yicI	5	σ 70
yjjJ	12	σ 70
ykgE	-41	σ 70
ykgE	25	σ 70
ymgG	-12	σ 70
ynaI	-11	σ 70
yqhC	40	σ 70
zapB	-13	σ 70
znuA	36	σ 70
znuCB	-11	σ 70
znuCB	-88	σ 70

Table B.2: Identification of the σ factors used for each RNAP binding site.

sites. However, they can also serve to sharpen the promoter's response to environmental cues. In the case of *ilvC*, IlvY is known to activate *ilvC* in the presence of inducer. However, we now see that it also represses the promoter in the absence of that inducer. The production of *ilvC* is known to increase by approximately a factor of 100 in the presence of inducer (Rhee, Senear, and Hatfield, 1998). The magnitude of the change is attributed to the cooperative binding of two IlvY binding sites, but the lowered expression of the promoter due to IlvY repression in the absence of inducer is also a factor.

Comparison of Reg-Seq results to regulonDB

B.9 Neural network fitting

Although neural network models are still in development, they are a more efficient fitting method than the Markov Chain Monte Carlo fitting methods used previously. Additionally this fits the results in k_bT units rather than arbitrary units as this method fits thermodynamic models. The training set for one neural network fit comprises DNA sequences from one promoter concatenated across all different experimental conditions. The categorical variable is meant to represent the experimental condition

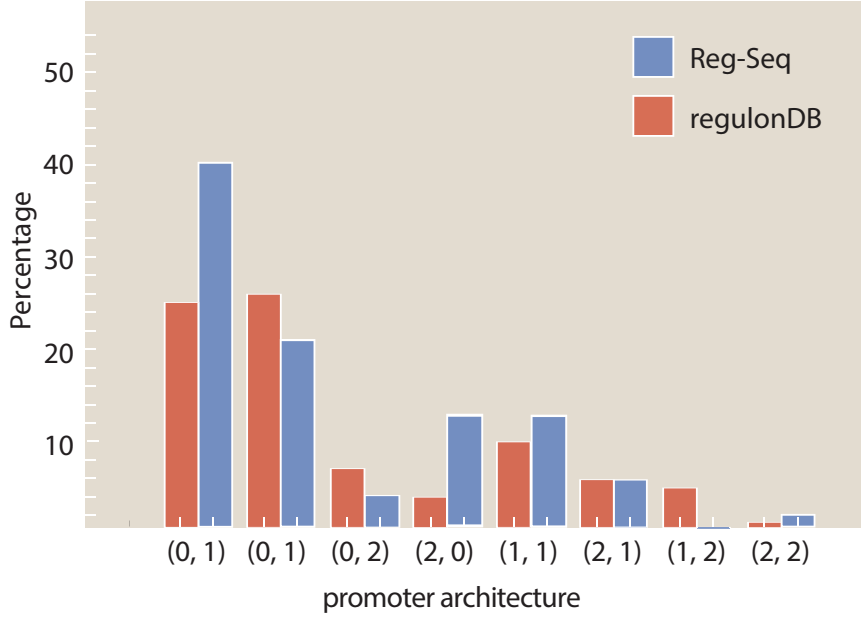


Figure B.9: Comparison of Reg-Seq architectures to RegulonDB. A comparison of the types of architectures found in RegulonDB (Santos-Zavaleta et al., 2019) to the architectures with newly discovered binding sites found in the Reg-Seq study.

(e.g. heat, M9 etc.). The training data are split as follows: 70% training, 20% validation, 10% test.

A schematic of the architecture of the neural network is shown in Fig. B.10 (schematic adapted from (Tareen and Kinney, 2019)). The sequence dependence of ΔG_C and ΔG_R is given by:

$$\Delta G_{TF} = \vec{\theta}_{TF} \cdot \vec{x}_{TF} + \vec{\mu}_{TF} \cdot \vec{z} + b_{TF} \quad (\text{B.21})$$

$$\Delta G_R = \vec{\theta}_R \cdot \vec{x}_R + \vec{\mu}_R \cdot \vec{z} + b_R. \quad (\text{B.22})$$

These Gibbs free energies are represented by the values of nodes in the first hidden layer. \vec{x} is a one-hot encoding of the input DNA sequence and \vec{z} is the condition categorical variable. One-hot encoding is a method to represent categorical variables. For example, if you were considering 3 growth conditions (M9, LB, Xylose), instead of labeling the categories M9=1, LB=2, Xylose=3, one-hot encoding labels M9= (1,0,0), LB=(0,1,0), and Xylose=(0,0,1). Without this encoding scheme, the "Xylose" growth condition would be weighted more than the "M9" growth condition

because the magnitude of the original label of "3" is larger than the magnitude of the "1" label of M9. With one-hot encoding, the magnitude of each label is the same.

μ represents the condition dependent part of the position weight matrix and b represents an overall bias/chemical potential. $\vec{\theta}$ represents the PWMs of the RNAP and the transcription factor. x_R represents the one-hot encoded sequence of the RNAP (similar for x_{TF}). The microstates of the thermodynamic model (see Fig. B.11), and equivalently the softmax activations of the second hidden layer, are given by

$$P_s = \frac{e^{-\Delta G_s}}{\sum_{s'} e^{-\Delta G_{s'}}} \quad (\text{B.23})$$

The nodes from the second hidden layer feed into a single, linearly activated, node representing transcription rate. A dense feed-forward network, with a Relu activated hidden layer and softmax activated output layer, maps transcription rate t to counts in bins. This network represents the error model $p(bin|t)$. The promoter activity t is given by Eq. B.24:

$$t = t_{sat} \frac{e^{-\Delta G_R} + e^{-\Delta G_R - \Delta G_{TF} - \Delta G_I}}{1 + e^{-\Delta G_{TF}} + e^{-\Delta G_R} + e^{-\Delta G_R - \Delta G_{TF} - \Delta G_I}} \quad (\text{B.24})$$

To fit the network, we minimize negative log-likelihood, given by Eq B.25:

$$\text{Loss Function} = -\frac{1}{\sum_{ij} c_{ij}} \sum_{i=1}^m \sum_{j=1}^N c_{ij} \log \left(P(\text{bin}_j | t(\vec{x})) \right). \quad (\text{B.25})$$

Here c_{ij} represents the counts of sequence i in bin j (N bins, and m sequences). Eq. B.25 represents log-Poisson loss, minimizing which is equivalent, in the large data limit, to maximizing mutual information $I[t, bin]$. We use stochastic gradient descent, in particular, the Adam optimizer, to back propagate losses.

For each promoter, the neural network model was fit 100 times; the two models with the lowest losses (each) on a held out test set were chosen to be the best models. The total procedure takes less than 15 minutes, which is a significant improvement from the several hours that fitting models with Markov Chain Monte Carlo can take. For future endeavors, where thousands of models will need to be fit, this is a crucial advance.

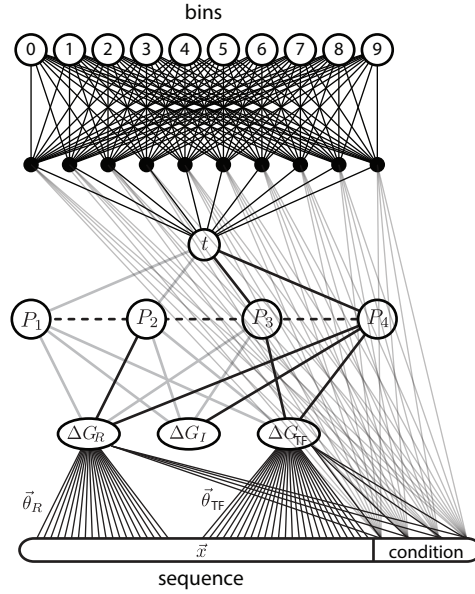


Figure B.10: Architecture of neural network used to fit data. \vec{x} represent a one-hot encoding of the input sequence. ‘condition’ is a categorical variable meant to represent the experimental condition of the experiment for each sequence. The condition variable feeds into in the energy nodes of the first hidden layer, and also to the dense non-linear sub-network mapping t to bins; this latter skipped connection has reduced opacity only to reduce visual clutter and does not represent any constraint on these skipped weights. Gray lines connecting first hidden layer weights to second hidden layer weights are fixed at 0. The weights linking nodes P_3 and P_4 to node t are constrained to have the same value is a diffeomorphic mode (Atwal, 2016).

B.10 Diffeomorphic modes

Diffeomorphic modes are parameters that we are unable to determine using maximization of mutual information because mutual information calculations are based on the rank order of model predictions, and changing these parameters does not change rank order of predictions. As one example there is a diffeomorphic mode that allows arbitrary scaling of energy binding matrices. This is why our matrices are, by default, reported using arbitrary units. However, more complex thermodynamic models can break diffeomorphic modes, as changing the previously undetermined parameters will now affect mutual information. A discussion of thermodynamic model fitting can be found in section 2.2 When we fit models to our data we maximize the mutual information between our predicted expression (from our model) and the actual binning. However, imagine the case where every cell’s fluorescence was doubled. The previously lowest expression cells would still be in the lowest

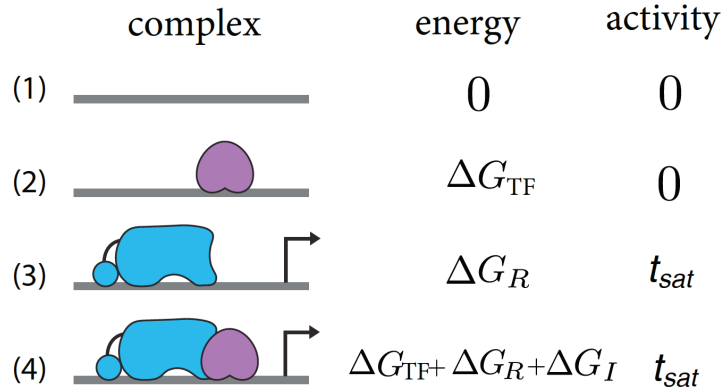


Figure B.11: The microstates of a one transcription factor promoter. We assume that any state with RNAP bound will transcribe at the rate given by "activity". The probabilities of each state can be calculated with a Boltzmann distribution.

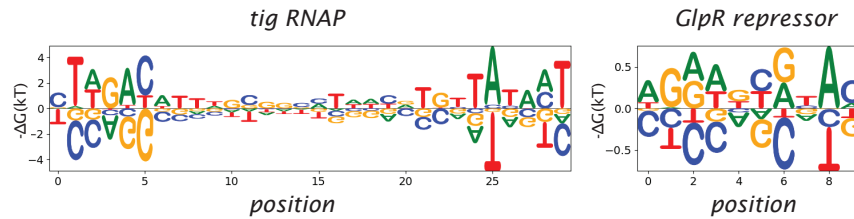


Figure B.12: A sequence logo of the *tff* which has one RNAP binding site and is repressed by GlpR. Fitting a thermodynamic model to the regulatory sequence allows absolute units (in k_bT units) to be fit to the energy matrix for the GlpR site.

bin, and the highest would still be in the highest bin, etc. Therefore this change would have no effect on the mutual information. Similarly, if we double our model's predicted expression for every cell, we would still predict they would be in the same bin, and mutual information would once again be the same. Mutual information is only sensitive to changes which affect the rank order of a cell's expression, and we are completely insensitive to any feature of our model which does not change rank order.

For constitutive expression, the expression is given by

$$\text{Expression} = \alpha \frac{\frac{P}{N_{ns}} e^{-\varepsilon_{Pd}}}{1 + \frac{P}{N_{ns}} e^{-\varepsilon_{Pd}}}. \quad (\text{B.26})$$

We need to use the value of expression as opposed to fold change because during the experiment we bin based on expression, not fold change, and the two measures are not equivalent. I will show they are not equivalent in the following section where we calculate the diffeomorphic modes for simple repression. One of the first thing we notice about this expression is that changing the value of α will not affect the rank order of any prediction (unless it is zero). This means we will not be able to determine α for constitutive expression. The second thing that pops out is that $\frac{P}{N_{ns}}$ could be absorbed into ε_P as an energy shift. We can see this transformation as follows

$$\frac{P}{N_{ns}} e^{-\varepsilon_{Pd}} = e^{-\varepsilon_{Pd} + \ln(\frac{P}{N_{ns}})}. \quad (\text{B.27})$$

We can then redefine

$$\varepsilon_P = -\varepsilon_{Pd} + \ln(\frac{P}{N_{ns}}), \quad (\text{B.28})$$

where $-\varepsilon_{Pd}$ is the product of a linear energy binding matrix and the DNA sequence in question, and θ is the energy binding matrix, and σ represents the sequence.

$$\varepsilon_{Pd} = \sigma^{mb} \theta_{mb} \quad (\text{B.29})$$

$$\varepsilon_P = \sigma^{mb} \theta'_{mb} \quad (\text{B.30})$$

$$\text{where } \theta'_{mb} = \theta_{mb} + \frac{\ln(\frac{P}{N_{ns}})}{\text{length of sequence}}. \quad (\text{B.31})$$

In this case we cannot distinguish between an additive shift to the entire energy binding matrix and the contribution from the concentration of binding protein $\ln(\frac{P}{N_{ns}})$. Therefore we can only ever determine one by setting a reference value for the other.

If we arbitrarily set α to 1, as its value is irrelevant, the expression equation can be rewritten as

$$\text{Expression} = \frac{1}{1 + R^{-1}}, \quad (\text{B.32})$$

where

$$R = e^{-\varepsilon_P/kT}. \quad (\text{B.33})$$

Cells with a higher value of R will always have a higher value of expression, and therefore rank order is preserved. This means that looking at only R is equivalent informationally to looking at the entire equation for expression. To be slightly more rigorous, R is an invertible function of expression and Kinney, 2008 proved this means they are informationally equivalent. If we look at only one binding site at a time, only the binding energy of that site will affect the expression. This is because all other features of the regulatory landscape can be treated as constants (albeit noisy constants).

This mutual information maximization method is very resilient to noise, and therefore only the energy of the single binding site needs to be taken into account in the model. It was proven by Kinney and Atwal, 2013 that in this one binding site case, the two diffeomorphic modes are an additive constant to the entire energy binding matrix, and a multiplicative scaling to the energy binding matrix. It was also shown that for a more complex system (for example looping), any possible diffeomorphic mode must also be a mode of one of the component transcription factors (i.e., for a system of 2 binding sites related by a thermodynamic model the ONLY 4 possible diffeomorphic modes are a multiplicative scaling of each binding matrix, and an additive shift to each). The equation for determining diffeomorphic modes of an arbitrary system is given by

$$g(\theta)\nabla_{\theta}R = h(R). \quad (\text{B.34})$$

The components of the energy binding matrix are given by θ_i . The left side of the expression gives a change in R when undergoing a transform g , the right side is an arbitrary function of R . To understand this equation lets look at a one base pair long energy matrix. In this case the binding matrix will be given by

$$\begin{array}{ll} \text{A} & \theta_A \\ \text{C} & \theta_C \\ \text{G} & \theta_G \\ \text{T} & \theta_T. \end{array}$$

We can define a transform to our energy matrix given by $\theta' \rightarrow \theta + \varepsilon g$, where g is

$\begin{pmatrix} g_A \\ g_C \\ g_G \\ g_T \end{pmatrix}$. Gene expression is then given by

$$\text{Expression} = \alpha \frac{1}{1 + e^\varepsilon}. \quad (\text{B.35})$$

Expression can equally be given by the value of ε because it is an invertible function

of expression. So we can define $R = -\varepsilon = -\theta \cdot \sigma$ where $\theta = \begin{pmatrix} \theta_A \\ \theta_C \\ \theta_G \\ \theta_T \end{pmatrix}$ and σ is a vector

which describes the base identity of the cell in question and is, for example equal to

$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ if the base is A and $\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ if it is a C. Then under the transformation of $g \rightarrow g'$, R

transforms from

$$R \rightarrow R' \quad (\text{B.36})$$

$$R' = R + \varepsilon g(\theta) \cdot \nabla_\theta R \quad (\text{B.37})$$

$$R' = R + \varepsilon g \cdot \begin{pmatrix} \partial_{\theta_A} R \\ \partial_{\theta_C} R \\ \partial_{\theta_G} R \\ \partial_{\theta_T} R \end{pmatrix}. \quad (\text{B.38})$$

If we define $b = \text{base}$, then

$$\partial_{\theta_i} R = \partial_{\theta_i} \varepsilon \quad (\text{B.39})$$

$$\partial_{\theta_i} \varepsilon = \delta_{bi} \quad (\text{B.40})$$

$$R' = R + \varepsilon \sigma \cdot g \quad (\text{B.41})$$

$$dR = \varepsilon \sigma \cdot g. \quad (\text{B.42})$$

Therefore

$$R' = R + \varepsilon \sigma \cdot g. \quad (\text{B.43})$$

If the differential change in R is a function of only the value of R, then there is no way that the rank order of the cells can change, and therefore the transformation

preserves information. If however, there is sequence dependence to the change in R (for example if $g_A = g_C = g_G = 0$ and $g_T = 1$) then rank order can change.

So to find the diffeomorphic modes of this system we need $dR = h(R)$, where h is any function of R .

$$g \cdot \sigma = h(R). \quad (\text{B.44})$$

For any choice of σ , this relation must hold true. One solution is that $g_A = g_C = g_G = g_T = \text{const}$ so that $h(R) = R + \text{const}$, corresponding to an additive energy shift. Another solution is that $h(R) = \beta R$ where β is some constant and $R = \sigma \cdot g$. Therefore the diffeomorphic mode is determined by

$$\sigma \cdot g = \beta \sigma \cdot \theta \quad (\text{B.45})$$

$$g = \beta \theta. \quad (\text{B.46})$$

This corresponds to a multiplicative energy shift. It was shown in Kinney and Atwal, 2013 that the only two proper diffeomorphic modes for a single binding site are this additive shift in binding energies and a multiplicative factor for the binding site. It is easy to see however that in our above simple example we can adjust any of the components of g up or down, as long as we do not change their rank order. This is because our possible values for the energy are discrete. Typically however, for energy matrices the size of a TF binding site, the possible energies will be nearly continuous, so in our analysis we will neglect these.

Justin calculated diffeomorphic modes for the single binding site case, as well as the simple activation case. I will do the same with simple repression, repressor looping, and the two overlapping promoter case.

B.11 Diffeomorphic Mode Calculations

I will define m as the base pair position, l as the length of the sequence, and b as the base pair identity A,C,G,T.

Simple Repression We can define Q as the statistical weight of repressor binding, namely $\frac{r}{N_{NS}} e^{-\beta \Delta \epsilon_r}$, where r refers to the number of repressors. Additionally we can define P as a similar statistical weight for RNAP, namely $\frac{p}{N_{NS}} e^{-\beta \Delta \epsilon_p}$, where p is the number of RNAP. Additionally, we will arbitrarily set $\alpha = 1$.

$$\text{Expression} = \frac{P}{1 + P + Q} \quad (\text{B.47})$$

$$\text{Expression} = \frac{1}{1 + R^{-1}} \quad (\text{B.48})$$

$$R = \frac{P}{1 + Q} \quad (\text{B.49})$$

R and expression are information equivalent. Any diffeomorphic mode of this system must be a diffeomorphic mode of one of the components. This is because we can look at a subset of data where the binding site sequences but one are constant. In this case the remaining binding site energy will be an invertible function of expression, and so the case is identical to the one binding site case. Adding additional data can not cause mutual information to cease depending on a particular parameter value. Therefore, although some of these modes will be broken in the more complicated cases, we only need to consider the diffeomorphic modes of each component.

Additive shift to the RNAP site

$$g(\theta_P) \nabla_{\theta_P} R = h(R) \quad (\text{B.50})$$

$$\text{where } g_{mb} = a_P = \text{constant} \quad (\text{B.51})$$

$$h(R) = g \nabla_{\theta_P} \frac{P}{1 + Q} \quad (\text{B.52})$$

$$P = \sigma^{mb} \theta_{P_{mb}} \quad (\text{B.53})$$

$$\nabla_{\theta_P} = \sigma^{mb} \partial_P \quad (\text{B.54})$$

$$h = \sigma^{mb} g_{mb} \partial_P \frac{P}{1 + Q} \quad (\text{B.55})$$

$$h = l a_P \partial_P \frac{P}{1 + Q} \quad (\text{B.56})$$

$$h = l a_P \frac{P}{1 + Q} \quad (\text{B.57})$$

$$h = l a_P R \quad (\text{B.58})$$

We see that $g(\theta) \nabla_{\theta} R$ is a function of R only, therefore an additive shift to the RNAP site energy is a diffeomorphic mode. We will not be able to determine the concentration of RNAP.

Multiplicative Shift to the RNAP site

$$g(\theta_P)\nabla_{\theta_P}R = h(R) \quad (\text{B.59})$$

$$\text{where } g_{mb} = b_P\theta_{P_{mb}} \quad (\text{B.60})$$

$$h = \sigma \cdot g\partial_P \frac{P}{1+Q} \quad (\text{B.61})$$

$$h = b_P P \partial_P \frac{P}{1+Q} \quad (\text{B.62})$$

$$h = b_P P \frac{P}{1+Q} \quad (\text{B.63})$$

$$h = b_P P R \quad (\text{B.64})$$

$g(\theta)\nabla_{\theta}R$ is a function not only of R but also of underlying sequence dependent elements. Therefore a multiplicative transform of P will alter mutual information and we will be able to precisely pin down the value by maximizing mutual information.

Additive Shift to the Repressor Site

$$g(\theta_Q)\nabla_{\theta_Q}R = h(R) \quad (\text{B.65})$$

$$\text{where } g_{mb} = a_Q = \text{constant} \quad (\text{B.66})$$

$$h = \sigma \cdot g\partial_Q \frac{P}{1+Q} \quad (\text{B.67})$$

$$h = la_Q \partial_Q \frac{P}{1+Q} \quad (\text{B.68})$$

$$h = la_Q \frac{R^2 Q}{P} \quad (\text{B.69})$$

$$(\text{B.70})$$

This equation can not be expressed as only a function of R , therefore an additive shift to the repressor site is not a diffeomorphic mode.

Multiplicative Shift to the Repressor Site

$$g(\theta_Q)\nabla_{\theta_Q}R = h(R) \quad (\text{B.71})$$

$$\text{where } g_{mb} = b_Q\theta_{Q_{mb}} \quad (\text{B.72})$$

$$h = \sigma \cdot g\partial_Q \frac{P}{1+Q} \quad (\text{B.73})$$

$$h = b_Q Q \partial_Q \frac{P}{1+Q} \quad (\text{B.74})$$

$$h = b_Q Q \frac{R^2 Q}{P} \quad (\text{B.75})$$

This equation can not be expressed as only a function of R , therefore a multiplicative shift to the repressor site is not a diffeomorphic mode. We can alternatively look at the expression for fold change.

$$\text{Fold-change} = \frac{1 + P}{1 + P + Q} \quad (\text{B.76})$$

$$\text{Fold-change} = \frac{1}{1 + R^{-1}} \quad (\text{B.77})$$

$$R = \frac{Q}{1 + P} \quad (\text{B.78})$$

The R for fold change has the repressor binding energy and RNAP binding energy exchanged compared to the R for expression. Therefore by symmetry the diffeomorphic mode for this system will be an additive shift to the repressor binding energy. Looking at fold change is not informationally equivalent to looking at expression, and since we bin based on raw expression, we need to look at expression.

B.12 Genes

Below will be a series of snippets on several genes to highlight from the RegSeq project, focusing on how features of the protein function and gene regulation can tell us more about the system. The amount of prior information varies wildly, with the genes newly found to be regulated by GlpR involved in the transcription, translation, or replication machinery of the cell, while many of those genes newly found to be regulated by FNR are members of the *y-ome* of *E. coli* and so have no knowledge of their function.

tff-rpsB-tsrf

rpsB and *tsrf* are both ribosome associated genes (a part of the 30S subunit and an elongation factor respectively). *rpsB* undergoes post transcriptional autoregulation. We found that there is also regulation at the transcriptional level by GlpR. We see that GlpR is affected by growth condition, particularly the presence of glucose. As ribosome amount is mainly dependent on growth rate it does make sense that some regulation would be dependent on glucose concentration. We see regulation at multiple levels, which should give a more precise level of control. As many genes must work together to produce the translational machinery it would be a very interesting future direction to see how preserved the mechanisms of gene regulation are across the different pieces of the machinery.

tig

Tig is a one of the chaperones which cooperate in the folding of newly synthesized cytosolic and secretory proteins (Keseler et al., 2013). As such, it is involved in the translation process, as *rpsB-tsrf* and *rhIE* all are. Similarly, it is also regulated by GlpR.

rhIE

RhIE is involved in ribosome maturation (Jain, 2008) and is regulated by GlpR. As such it is also a part of the group of translation related genes regulated by *GlpR*. As GlpR will be induced by the presence of glucose or absence of glycerol, this is most likely a feature of how both translation machinery and the presence of glucose are correlated with growth rate. *rhIE* was known to have highly differential regulation from Schmidt, Kochanowski, Vedelaar, Ahrné, et al., 2016, and we see in our data that it repressed in glucose containing conditions compared to other growth conditions.

maoP

MaoP contributes to the positioning of the Ori macrodomain (Valens, Thiel, and Boccard, 2016), which is crucial for positioning the origin of replication. Additionally, it is involved in stress induced mutagenesis (Al Mamun et al., 2012). MaoP is an important protein for replication of DNA. We have found two new transcription factors binding sites (GlpR, and PhoP) and with the addition of HdfR, for which we have confirmed the binding location.

rapA

RapA interacts directly with RNAP and is involved in recycling stalled RNAP. RapA is repressed by GlpR and as shown in Fig. 4.6 (A), and when GlpR is active, the most active RNAP site is changed to an upstream site at -105 bp from the original TSS. The leakiness of transcription from *rapA* will be higher than it would be otherwise.

ybjX

Lipid A, is the hydrophobic moiety of lipopolysaccharide (LPS), a glucosamine-containing saccharolipid that constitutes the outer layer of the outer membrane of most gram-negative bacteria (Raetz et al., 2007). Lipid A synthesis is catalyzed by

WaaA, HtrB, MsbB and PagP (regulated by PhoP). *htrB* or *msbB* knockouts have been shown to cause growth defects in both *S. enterica* serovar *Typhimurium* (from now on *S. typhi*) (Murray et al., 2001). The lipid A moiety of LPS is detected by the TLR4/MD2 receptor of the mammalian innate immune system, and modifications of lipid A can protect bacteria from antibiotics (Raetz et al., 2007).

YbjX is a relatively understudied protein in *E. coli*, and there is little information available in knowledgebases like Ecocyc and RegulonDB. It is known however, mutations in the YbjX homolog in *S. typhi*, SomA, caused partial suppression of the growth defects of an $\Delta msbB$ strain (Murray et al., 2001). In this sense, the *somA* mutation partially compensates the loss of function mutation of the *msbB* gene in *S. typhi*.

Interestingly, using data from Price *et al.* we found that $\Delta ybjX$ mutants had a fitness advantage when grown on 1mg/mL bacitracin- an antibiotic targeting cell wall and peptidoglycan biosynthesis. However, the $\Delta ybjX$ mutant also caused a fitness penalty when grown on 0.001 mg/ml doxycycline-an antibiotic that inhibits protein synthesis by binding to the 30S ribosomal subunit, and in 0.006 mg/ml nalidixic acid-which blocks DNA replication (Price et al., 2018). Overall these results suggest that a *ybjX* loss-of-function mutation confers a positive effect on cell wall defense by an unknown mechanism.

Using RegSeq we found that the *ybjX* promoter is controlled by an activator. We found that PhoP binds to the *ybjX* promoter using mass spectrometry. The PhoP regulon comprises genes that act on the adaptation to low Mg^{+2} , acid resistance (Zwir et al., 2012), and antibiotic efflux pumps AcrAB and TolC. PhoP is also related to virulence factors in *S. typhi* (Monsieus et al., 2005). In this sense, YbjX is a member of the antibiotic resistance and stress response regulon in *E. coli* K-12 via PhoP regulation (Monsieus et al., 2005). Further studies need to be conducted to elucidate the mechanism of action *ybjX* on cell wall synthesis and the overall pleiotropic effects of this protein.

B.13 Construction of sequence logos

With our position weight matrices in hand we can now construct sequence logos by calculating the average information content at each position along the binding site. With our four letter alphabet there is a maximum amount of information of 2 bits ($\log_2 4 = 2$ bits) at each position *i*. The information content will be zero at a position when the nucleotide frequencies match the genomic background, and will have a

maximum of 2 bits only if a specific nucleotide is completely conserved. The total information content at position i is determined through calculation of the Shannon entropy, and is given by

$$I_t = \sum_{j=A}^T p_{ij} \cdot \log_2 \frac{p_{ij}}{b_i}, \quad (\text{B.79})$$

were the terms in the summation $\log_2 \frac{p_{ij}}{b_i}$ are the terms in the position weight matrix (Schneider et al., 1986; Stormo, 2000). The total information content contained in the position weight matrix is then the sum of information content across the length of the binding site.

To construct a sequence logo, the height of each letter at each position i is determined by

$$\text{logo height}_{ij} = p_{ij} \cdot I_i, \quad (\text{B.80})$$

where h is the height displayed on the sequence logo and is in the units of bits. The relative height of the bases A , C , G , and T are displayed according to their relative probabilities scaled by information content (Schneider et al., 1986). We construct sequence logos using WebLogo (Crooks et al., 2004) and custom code from Justin Kinney available on the GitHub repo for the Sort-Seq project (https://github.com/RPGroup-PBoC/sortseq_belliveau).

Comparison of Sort-Seq sequence logos.

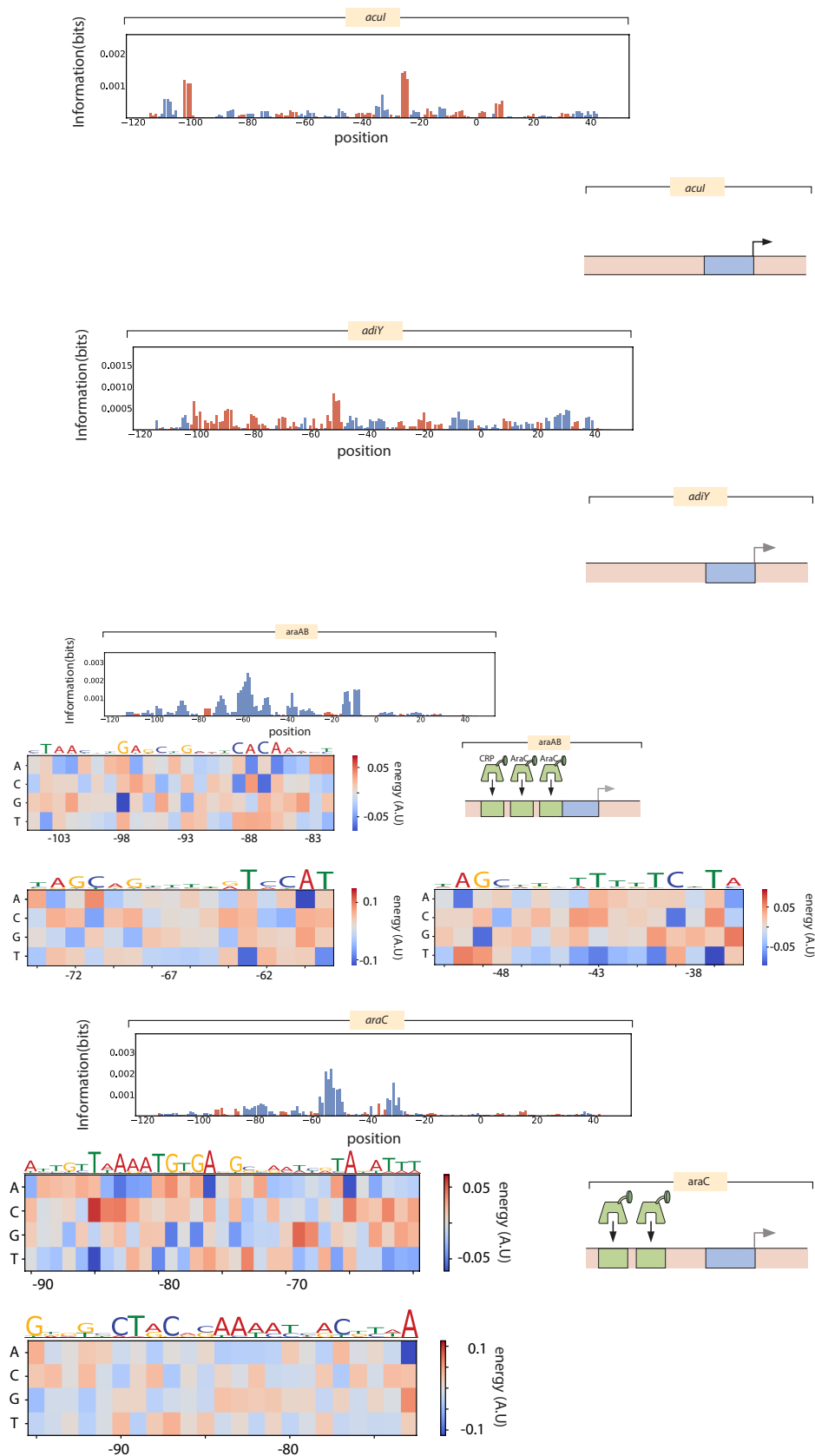
For the various annotated binding sites identified in this work we used our Sort-Seq data to generate energy matrices. We have also found it easy to visualize the sequence preference as sequence logos, and we can compare our generated sequence logos to those created by studying several known sites in the *E. coli* genome. In Fig. 3.4 we show a comparison of logos found via Sort-Seq for transcription factors with three or more known genomic binding sites, with agreement more apparent when genomic binding site logos are based on a larger number of known sequences. We also report the Pearson correlation coefficient between the position weight matrices from the Sort-Seq inference and the genomic alignment. To compare the two position weight matrices we first apply gauge fixing to each matrix in a similar manner as our energy matrix. Each column is set to have a mean energy of zero and the matrix norm (or

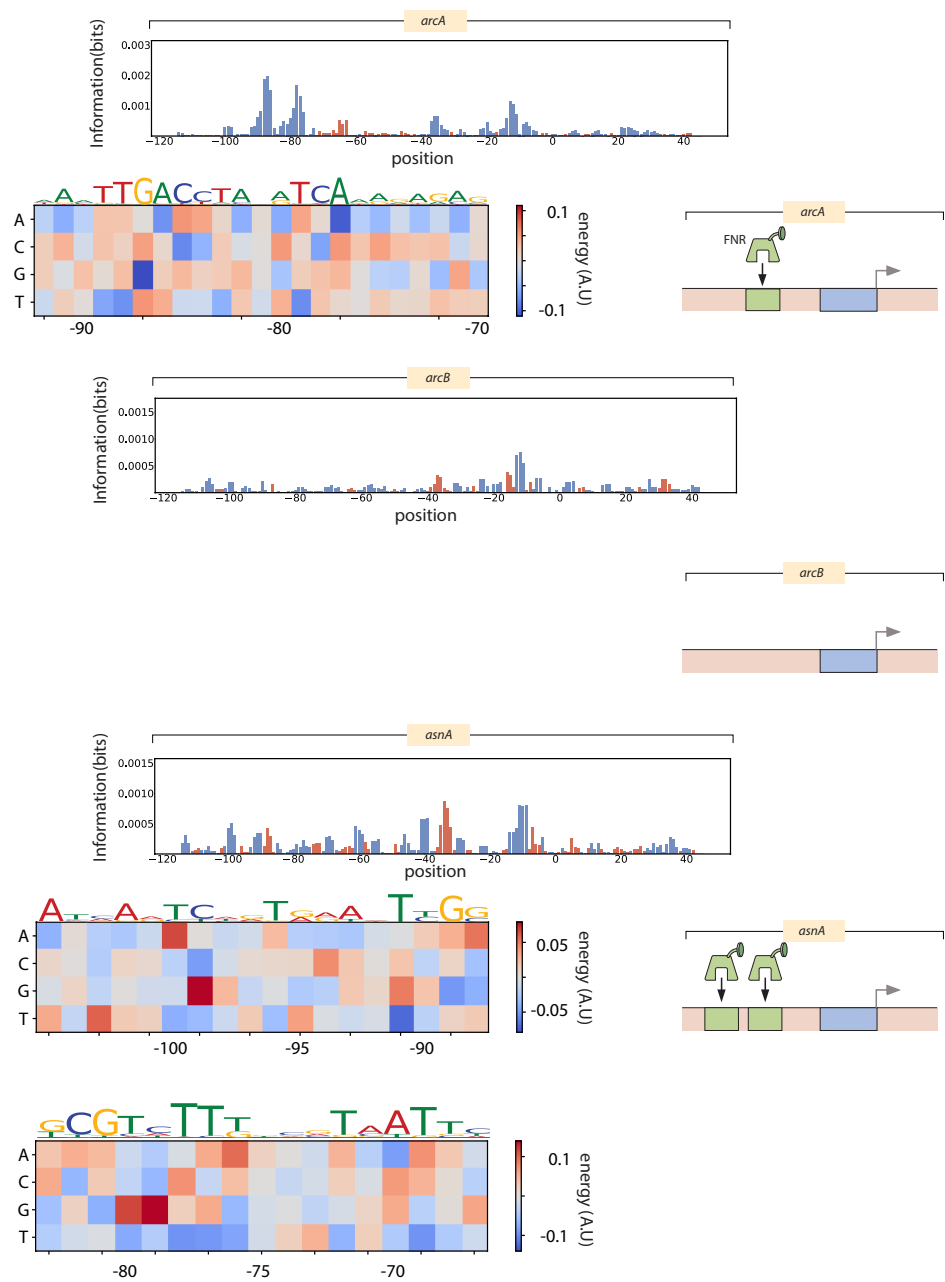
inner product) is normalized to have value one. Under this constraint, the Pearson correlation coefficient is simply given by the summed product of matrix entries,

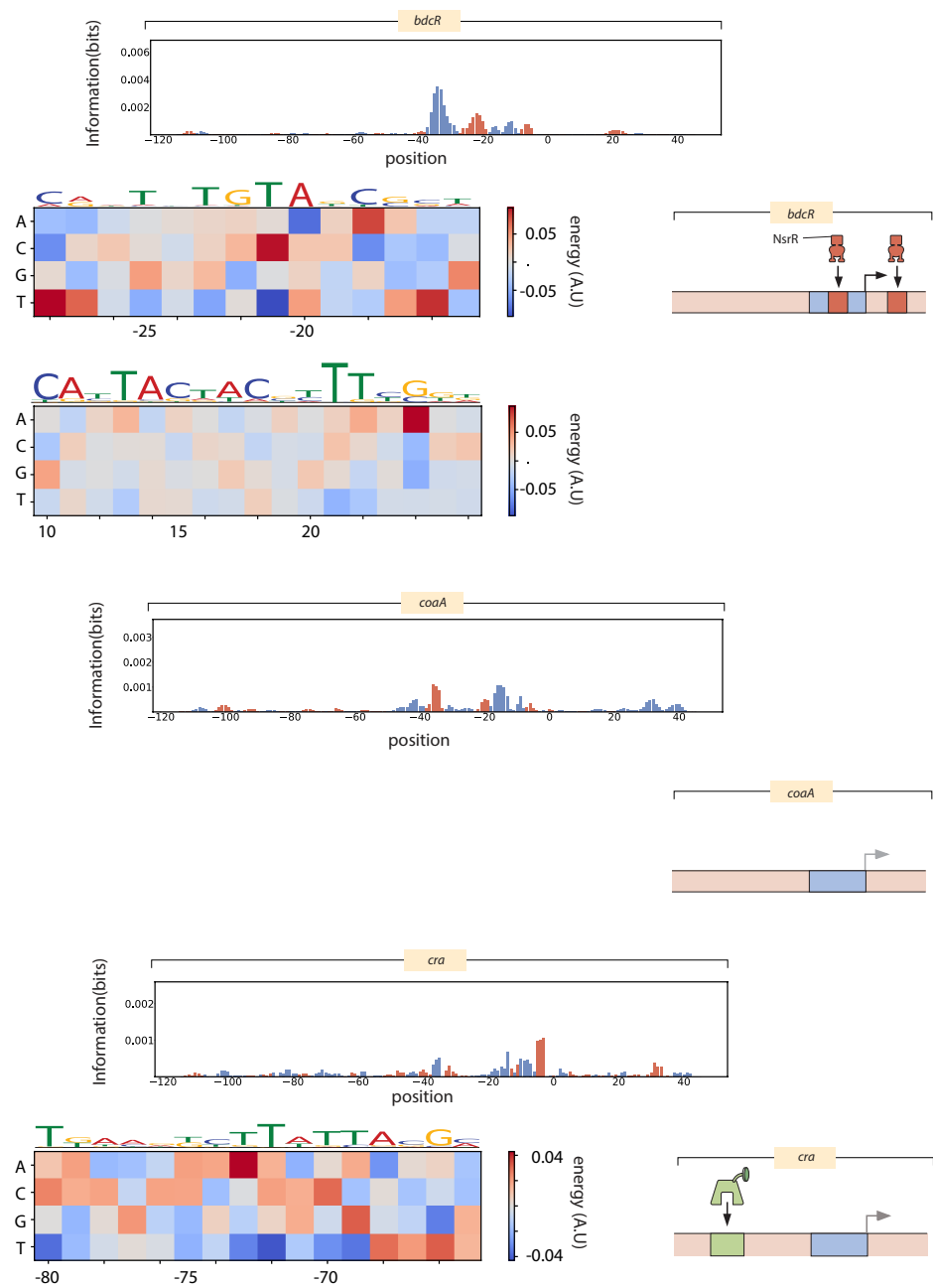
$$r = \sum_{i=1}^L \sum_{j=A}^T PWM'_{X,i,j} PWM'_{Y,i,j}, \quad (\text{B.81})$$

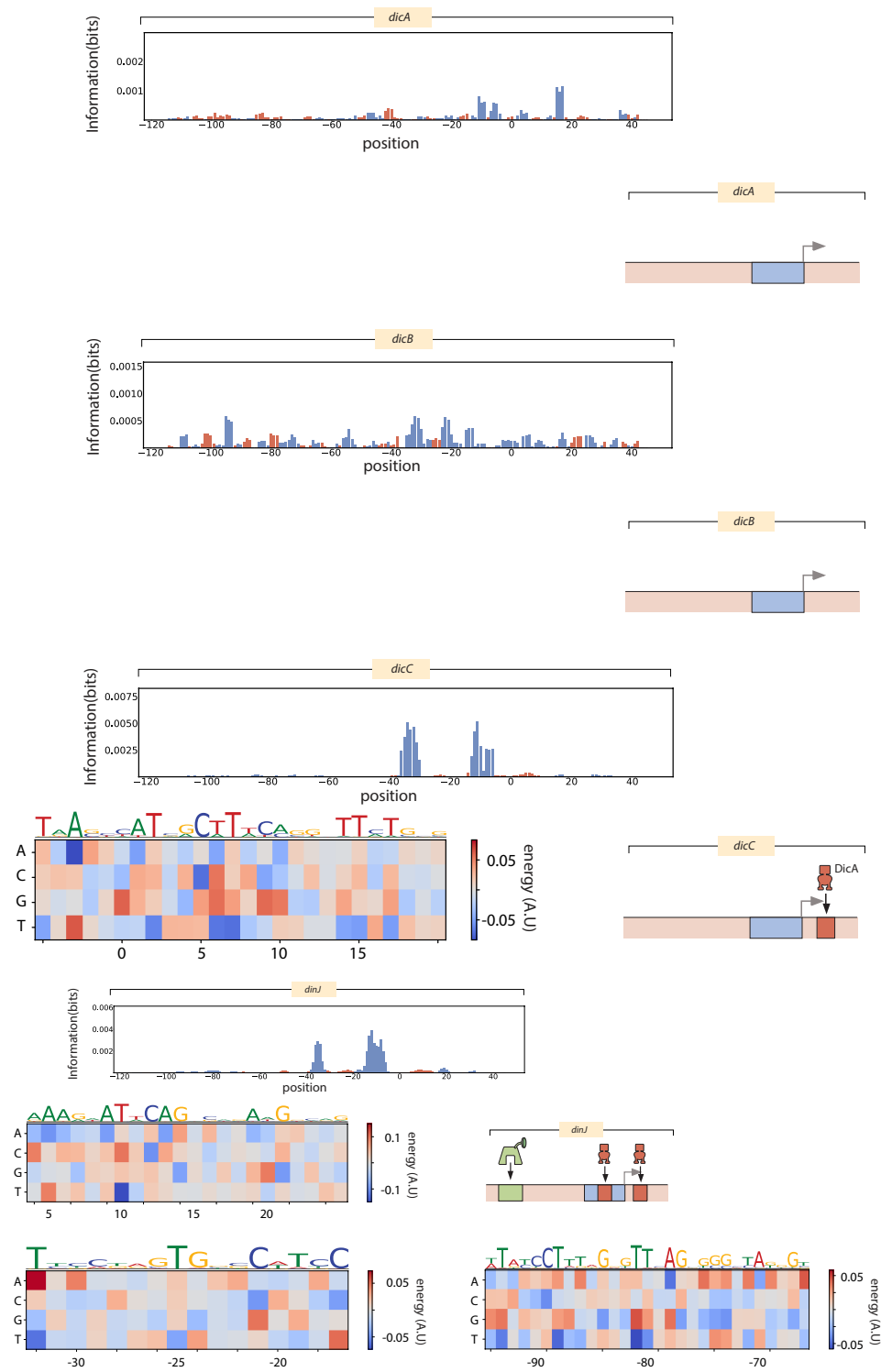
where X and Y refer to the two different PWM being compared. We do a similar comparison between models generated via Sort-Seq and Reg-Seq in Fig. B.2.

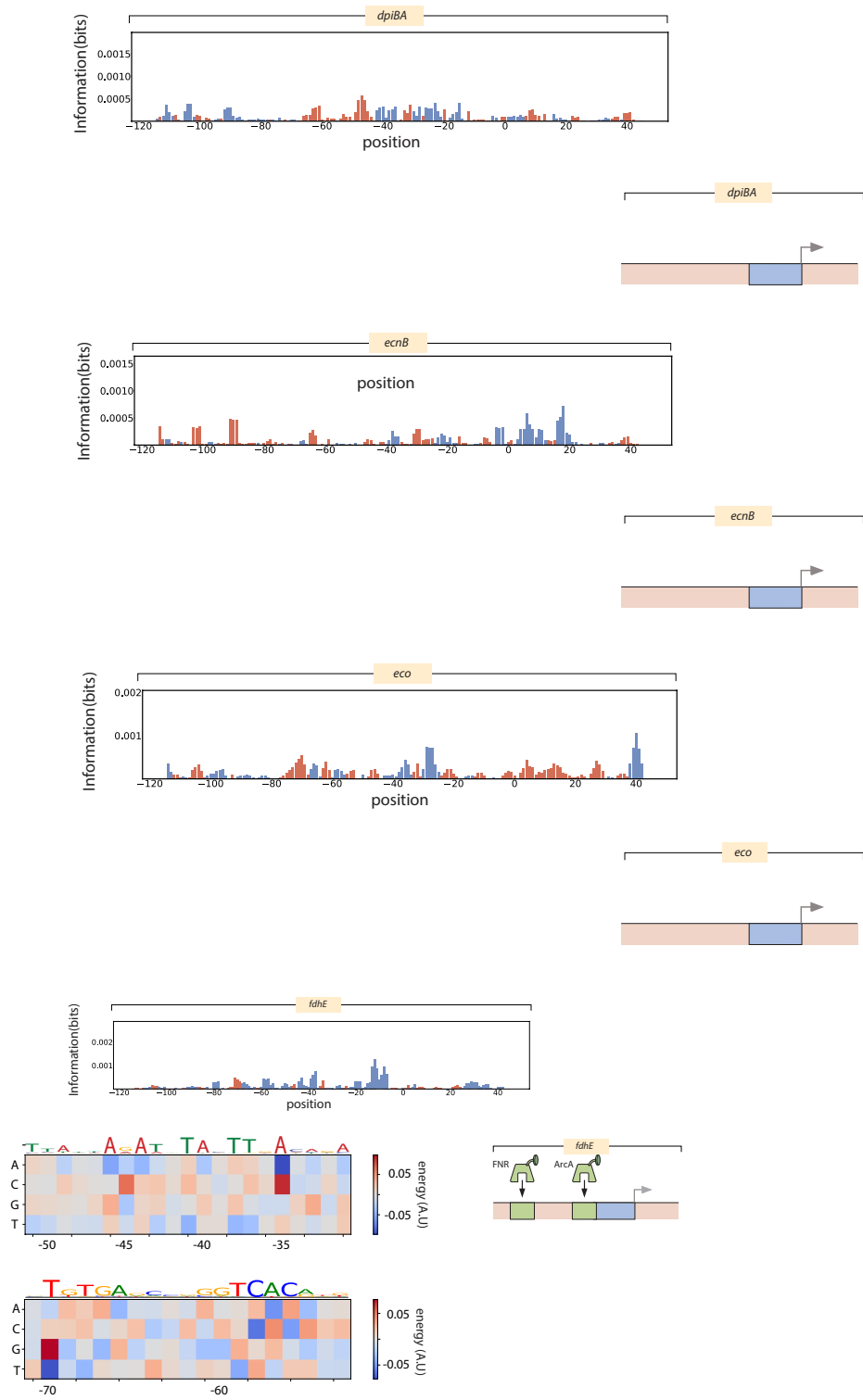
Reg-Seq Data

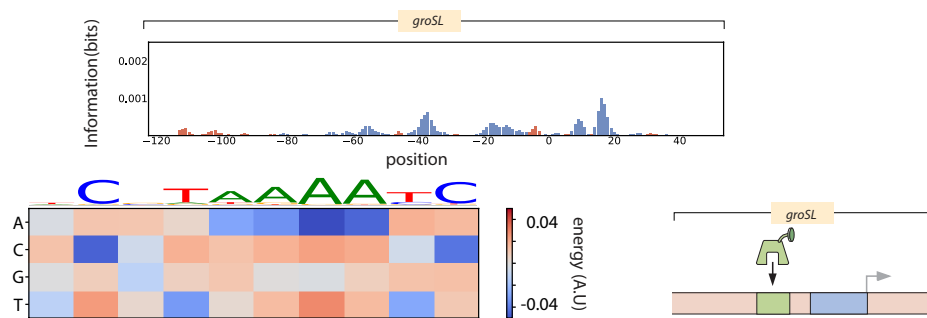
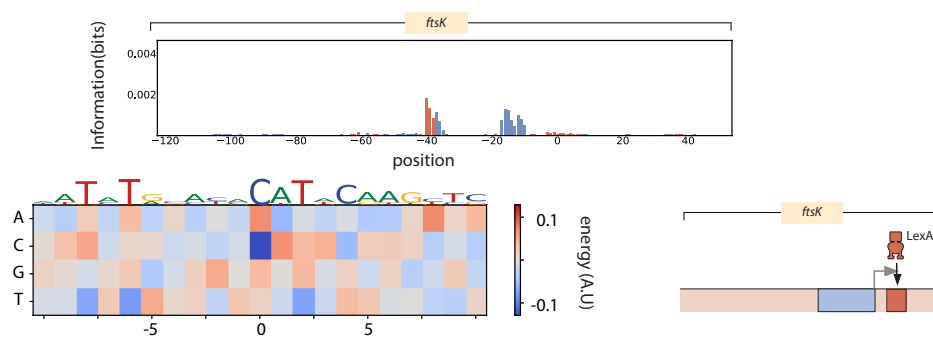
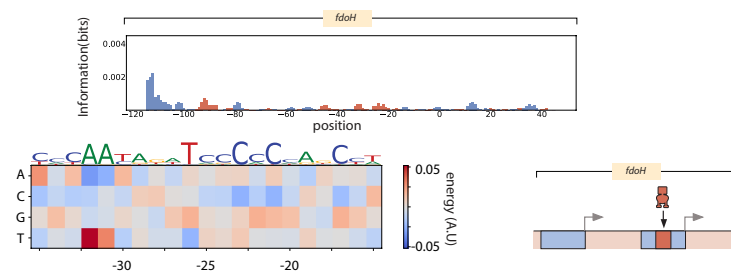


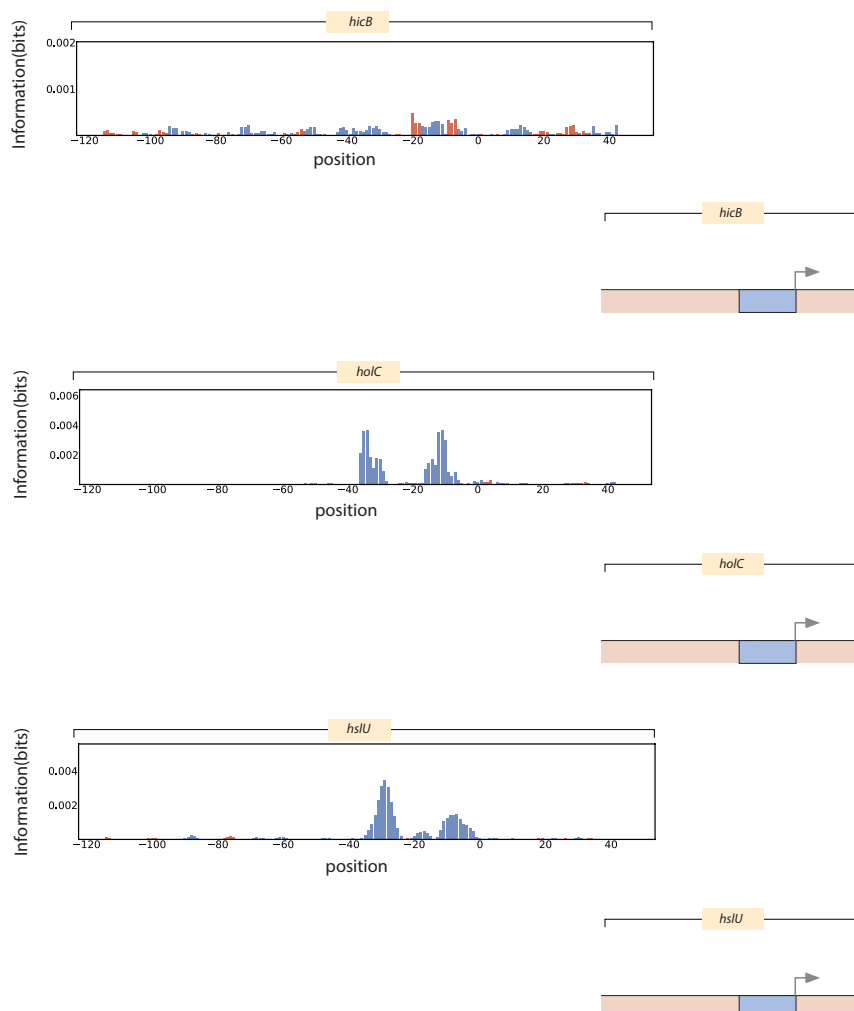


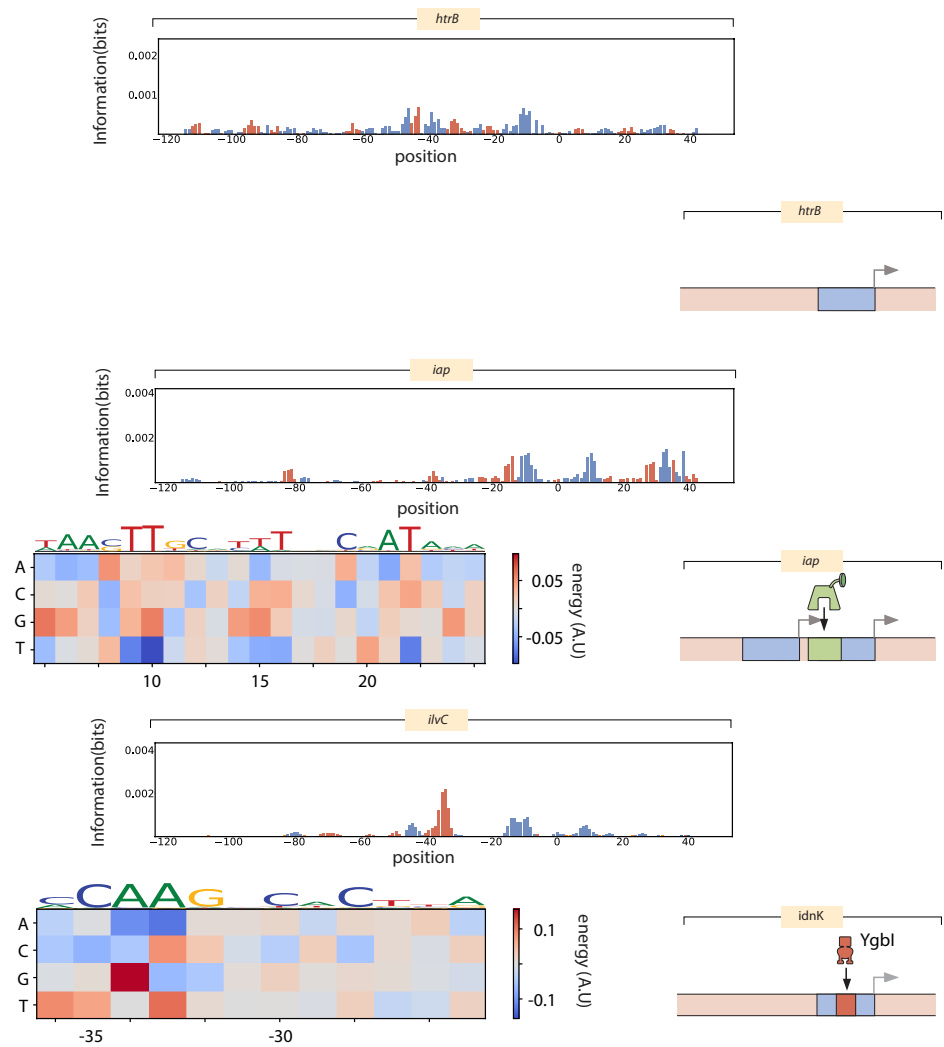


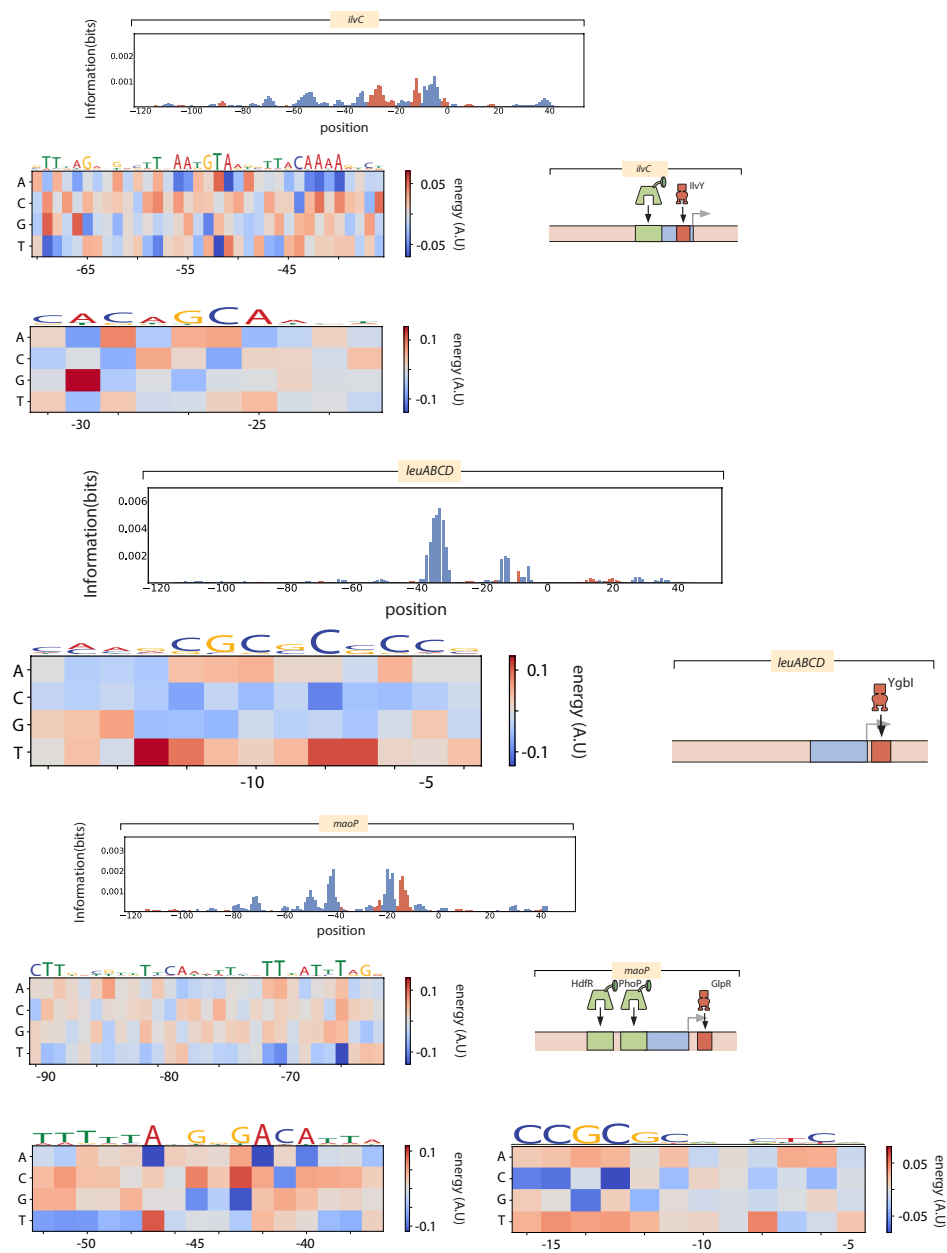


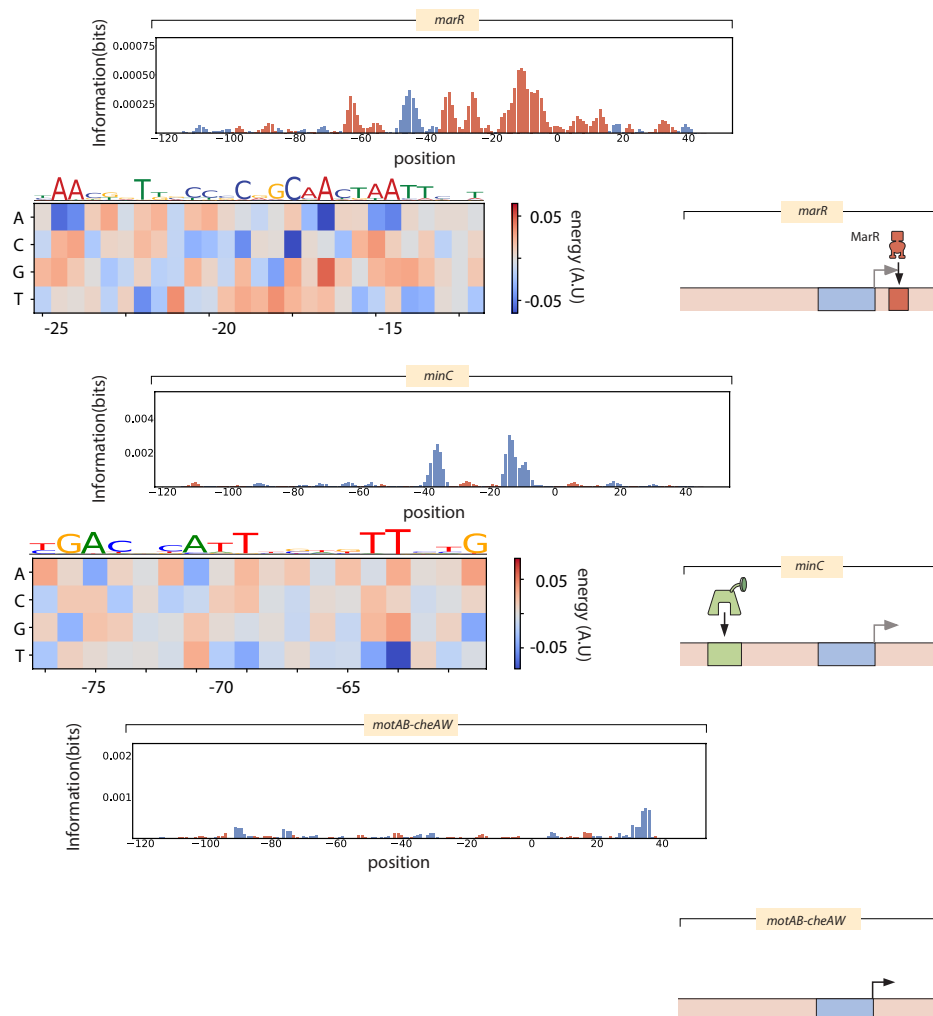


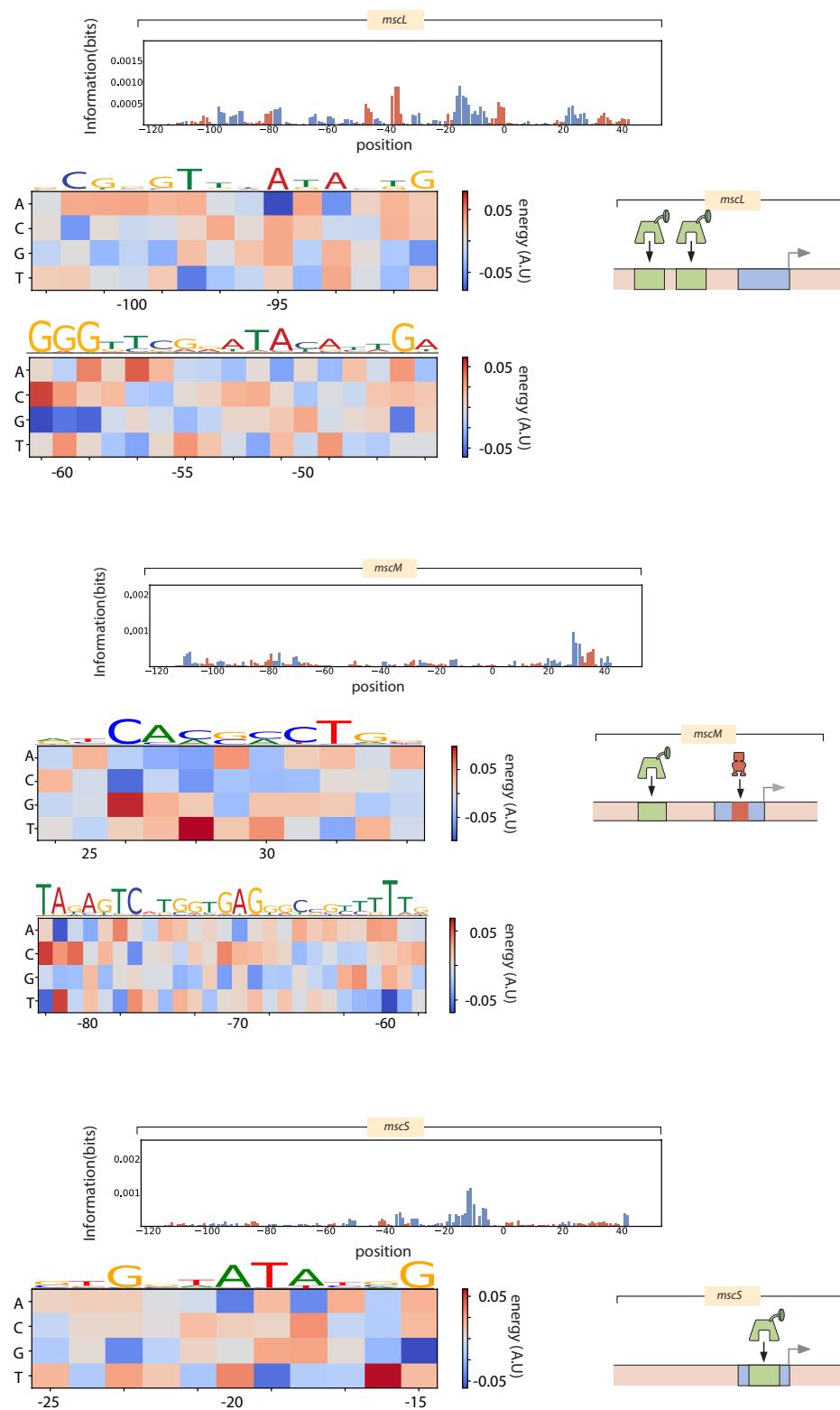


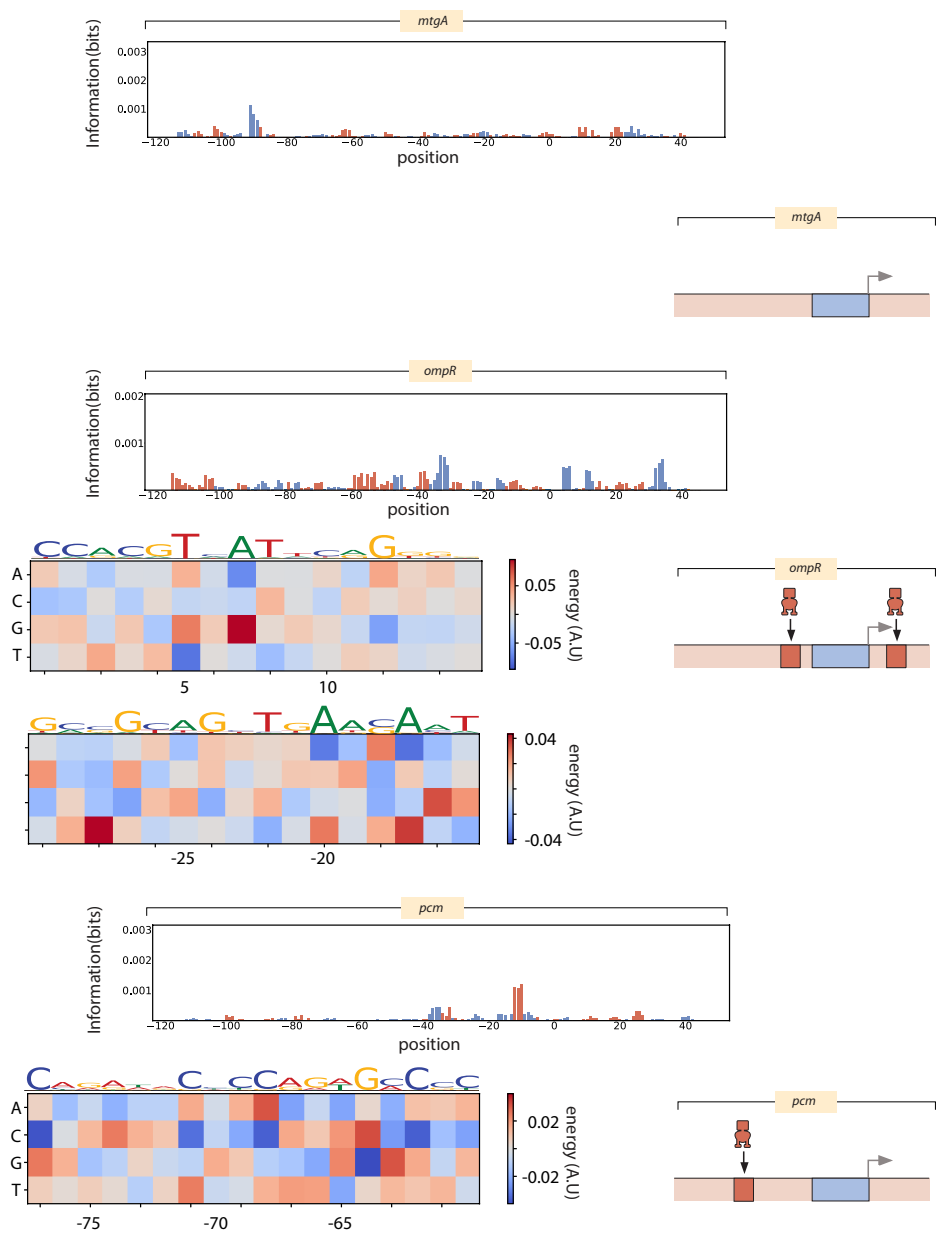


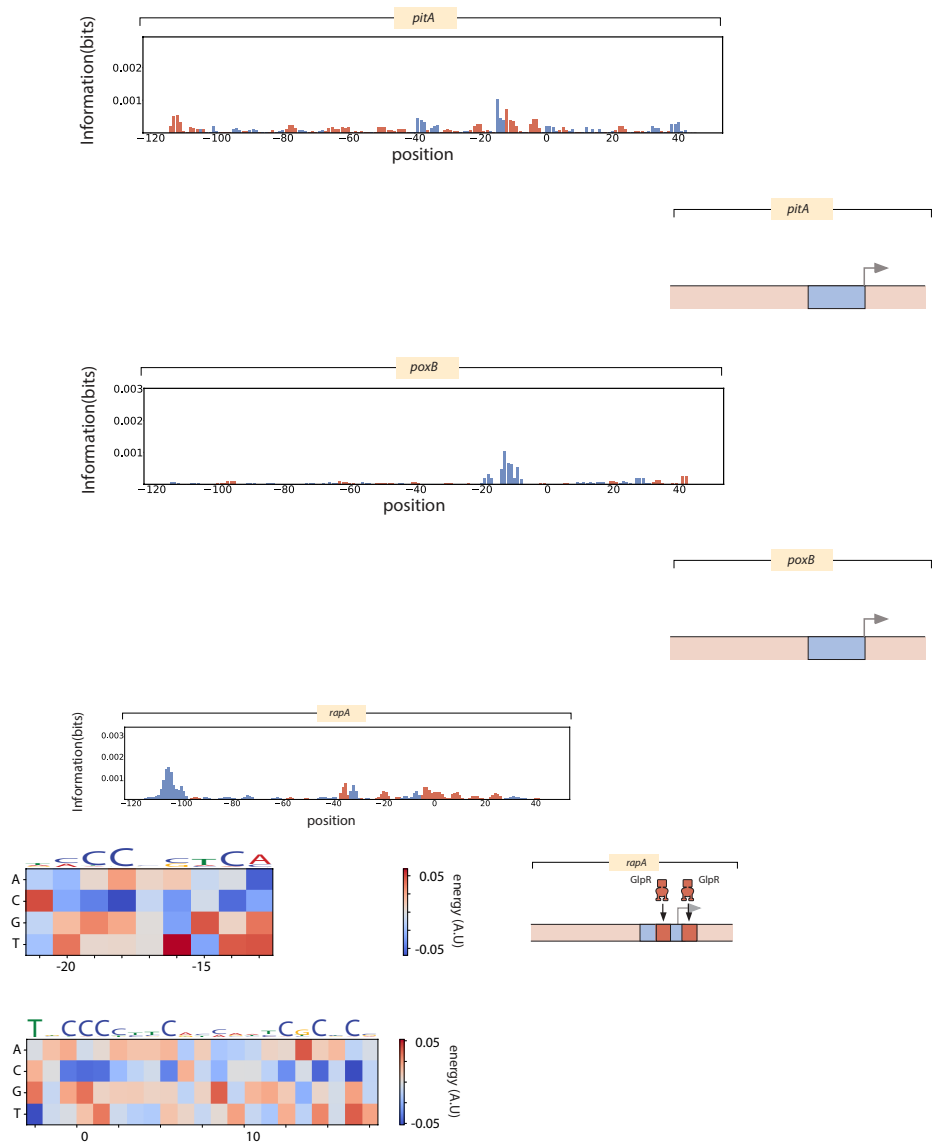


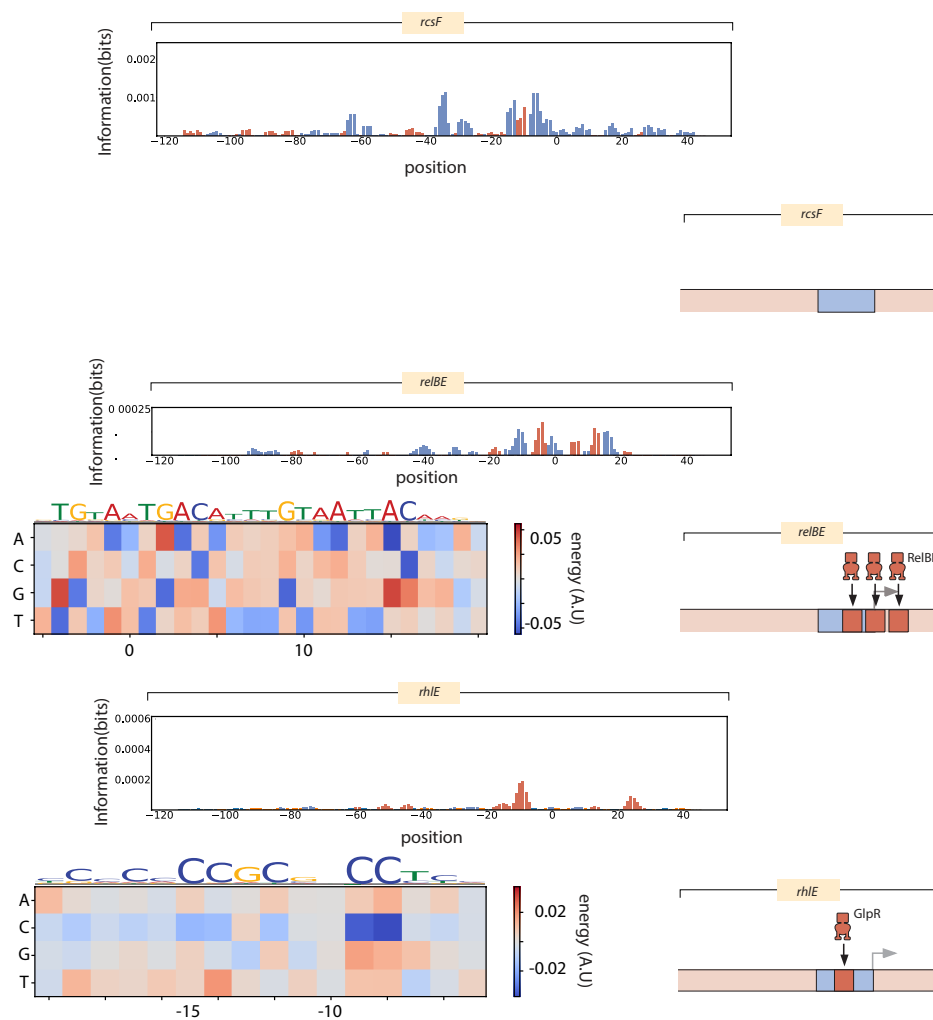


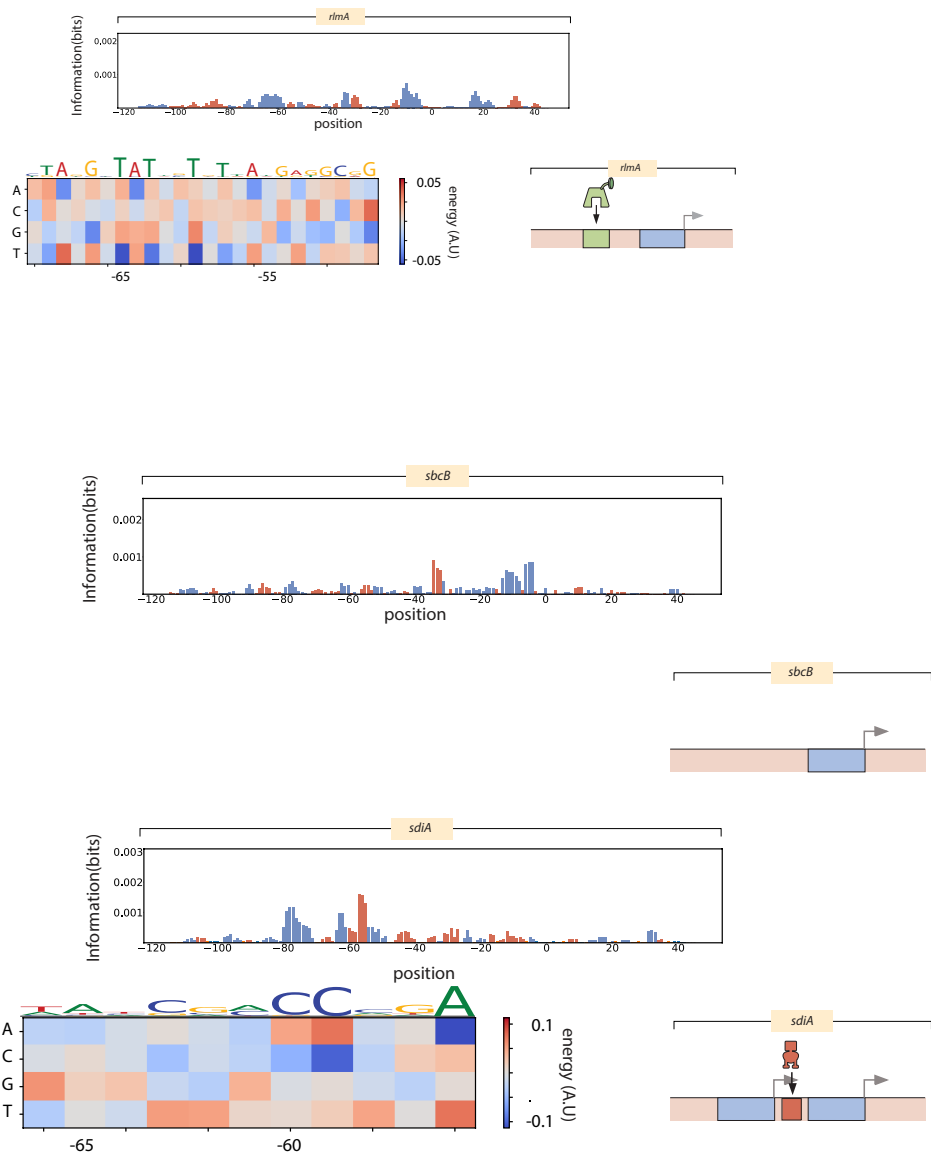


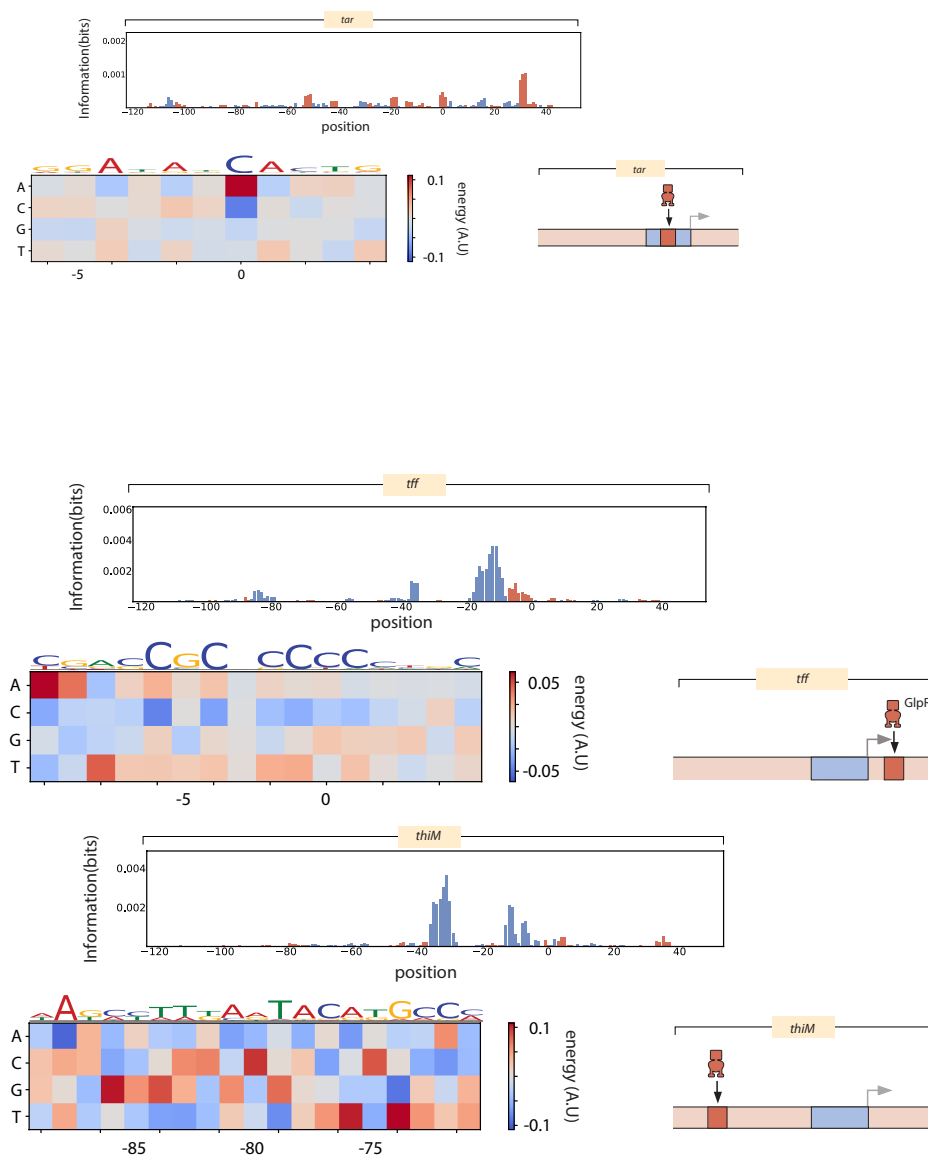


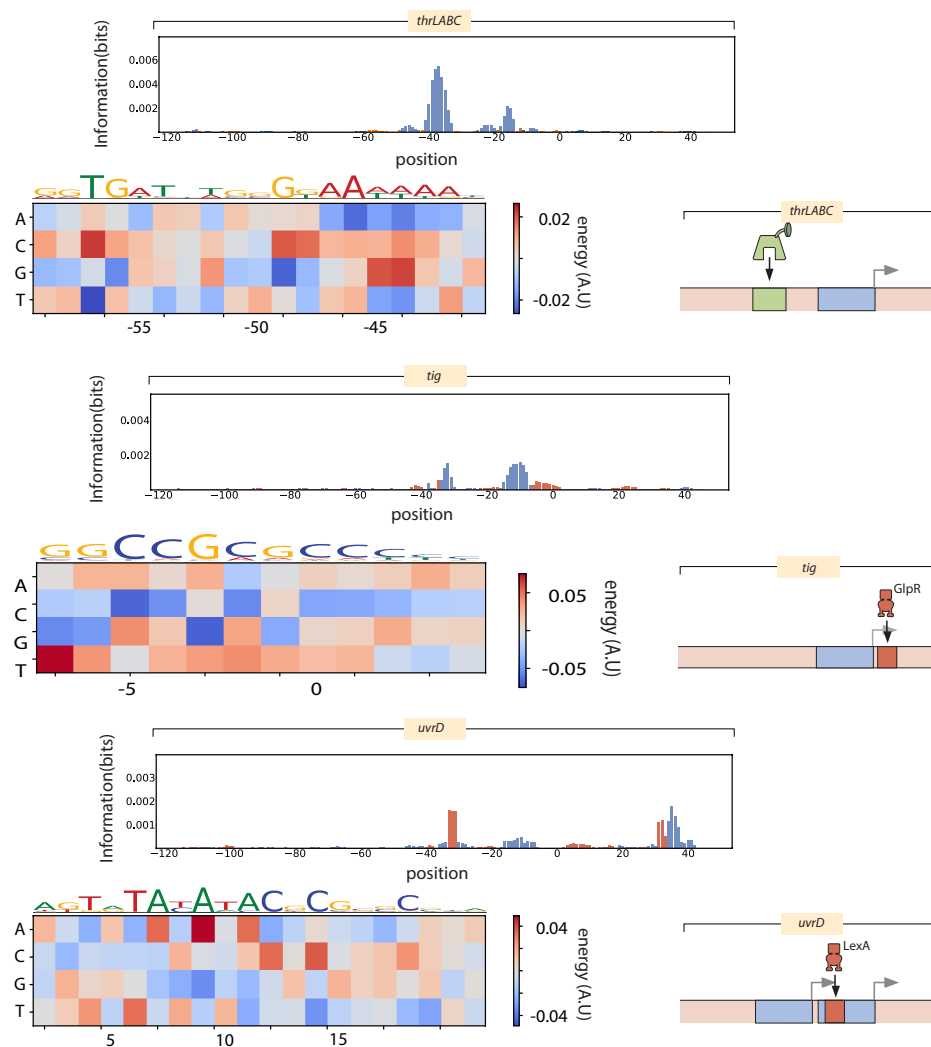


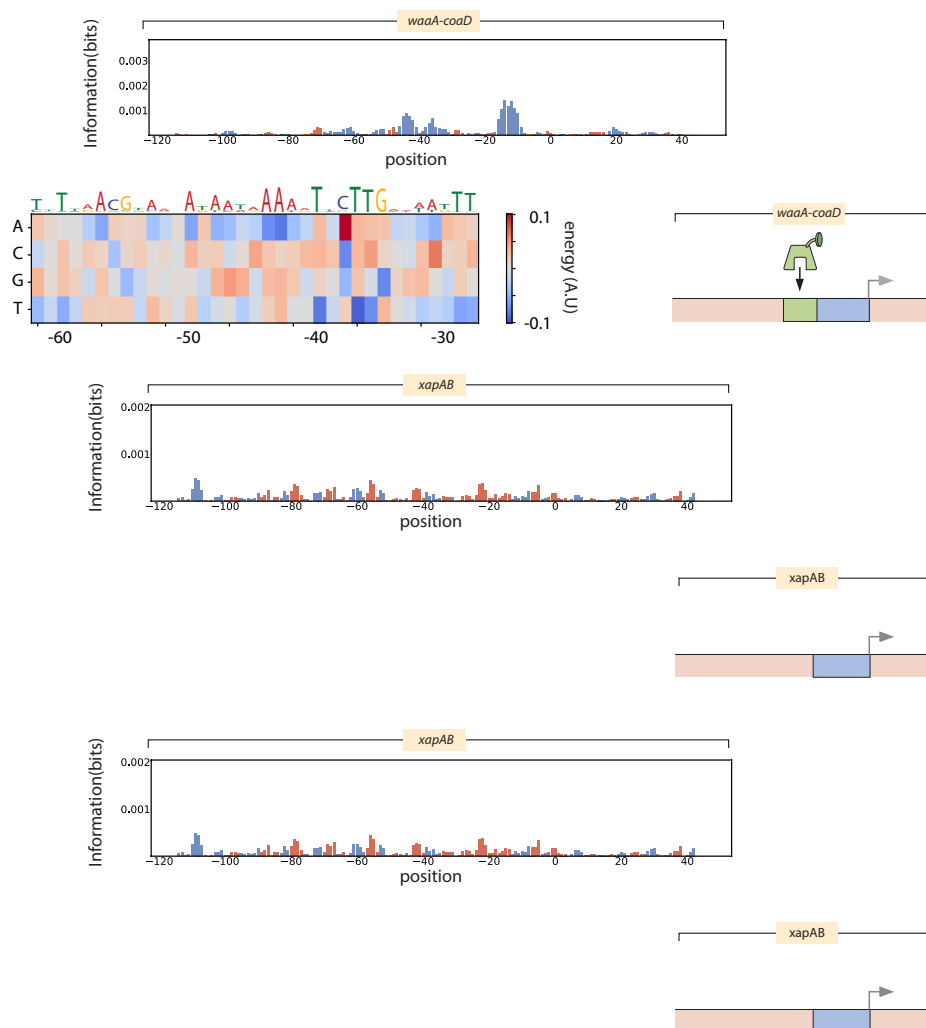


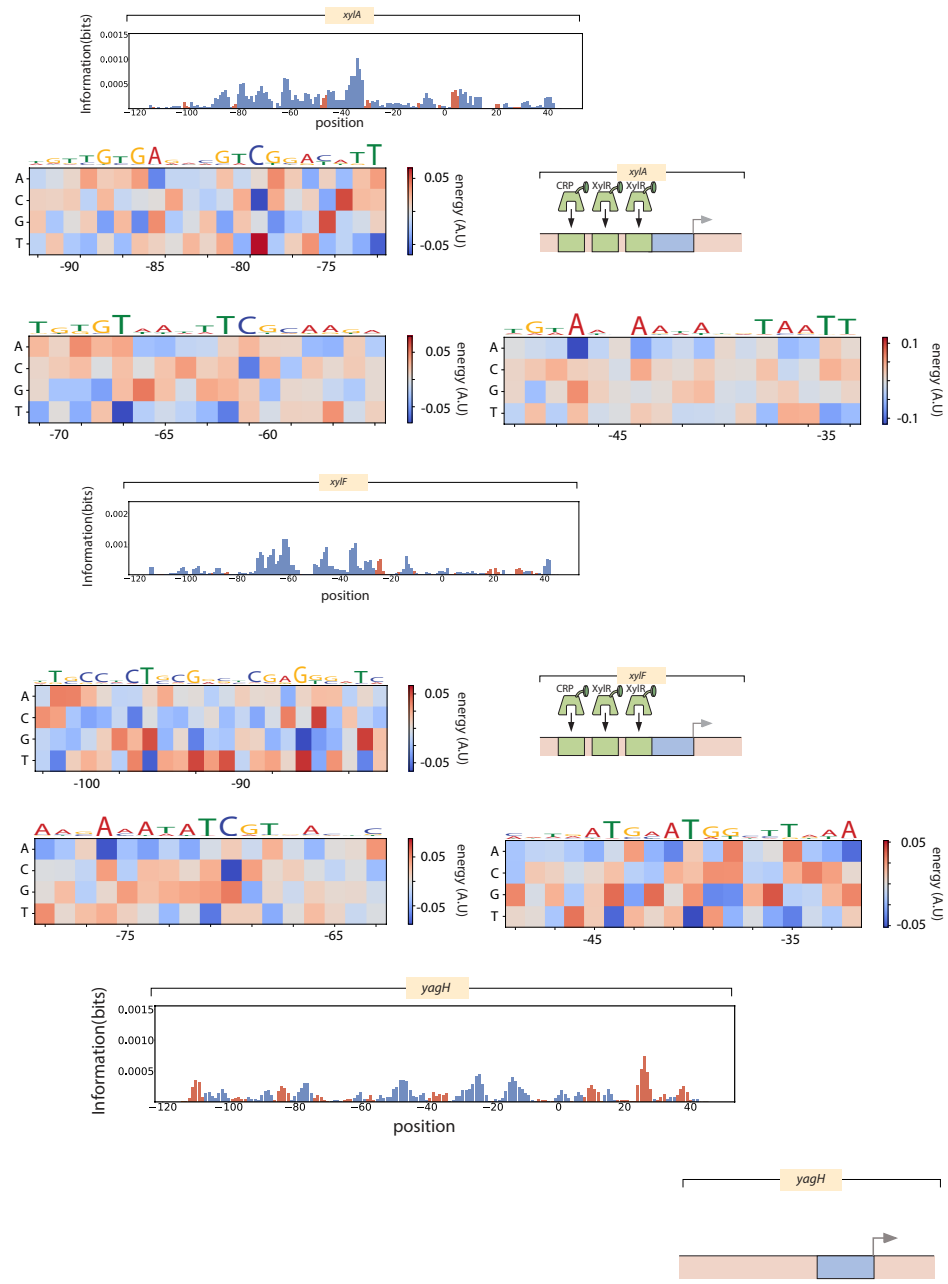


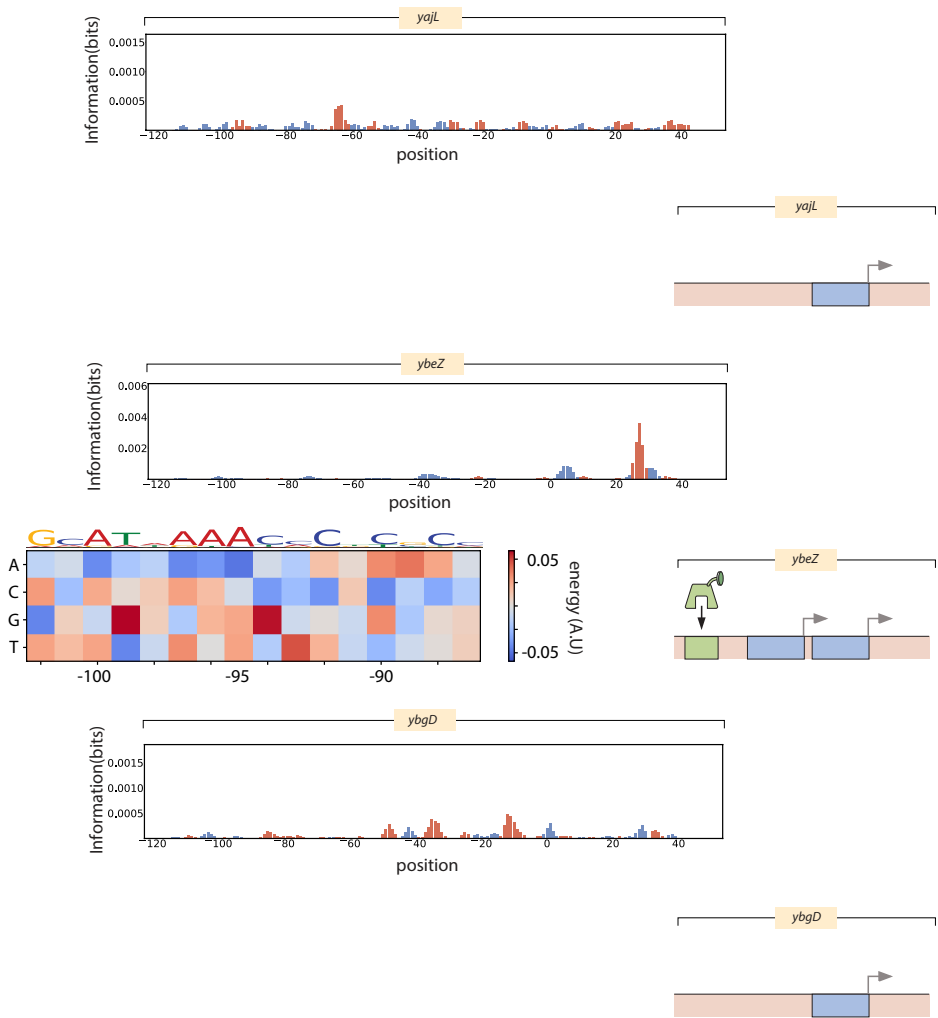


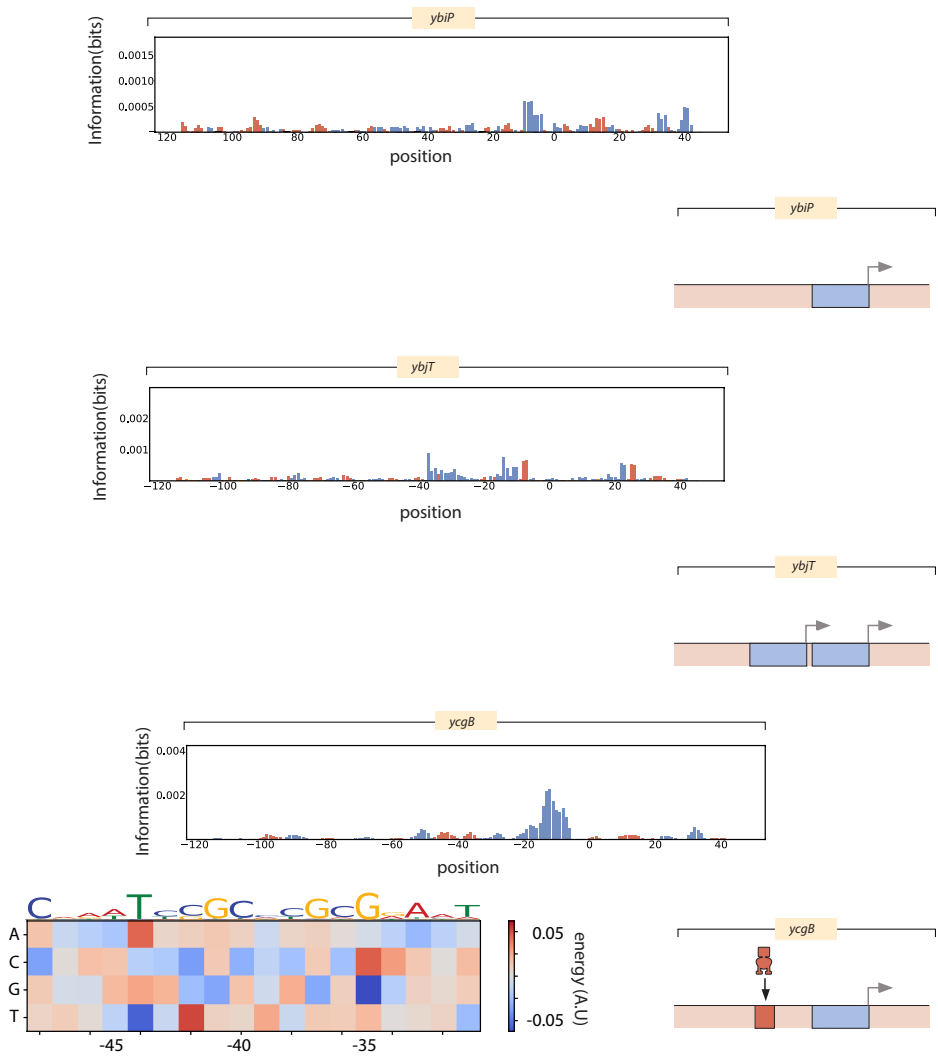


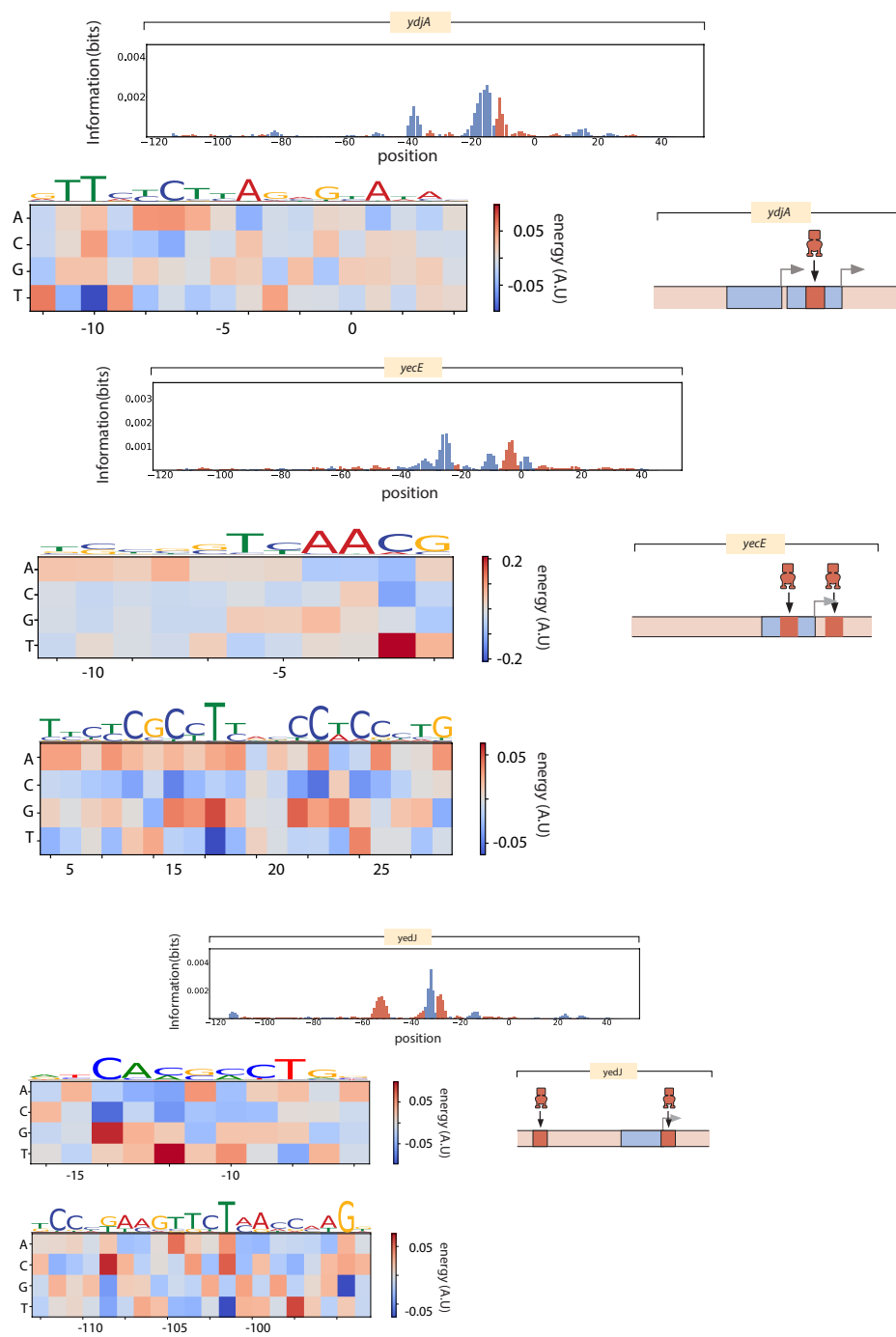


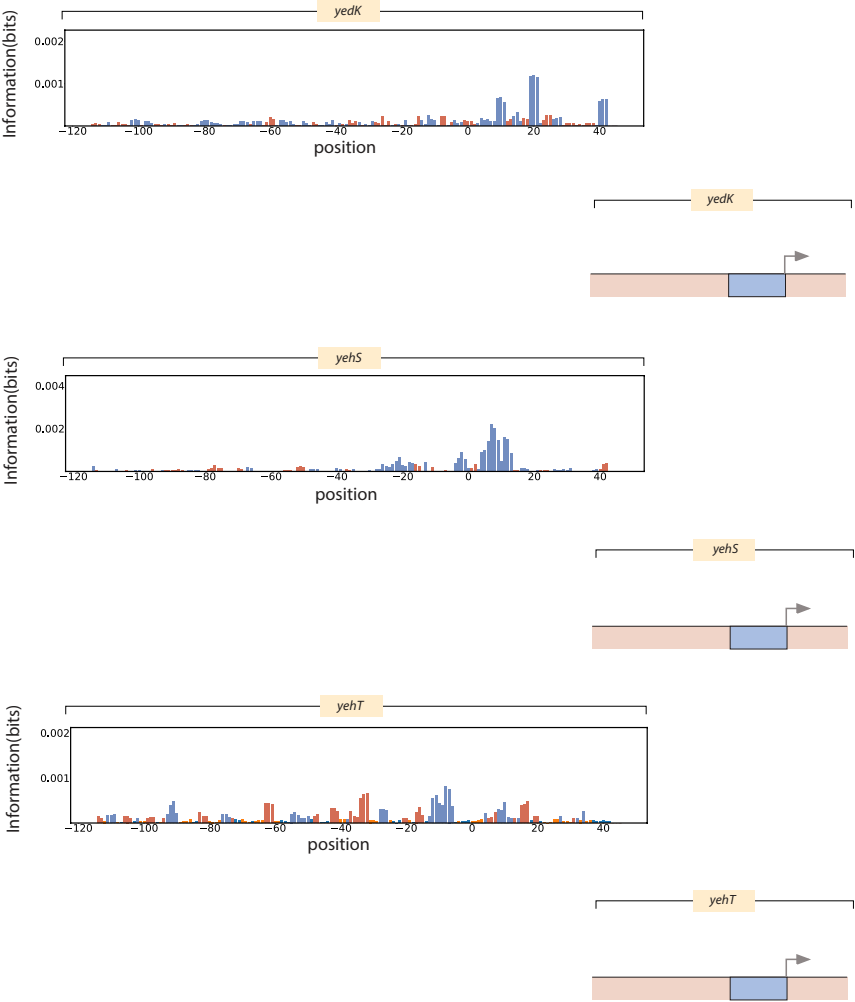


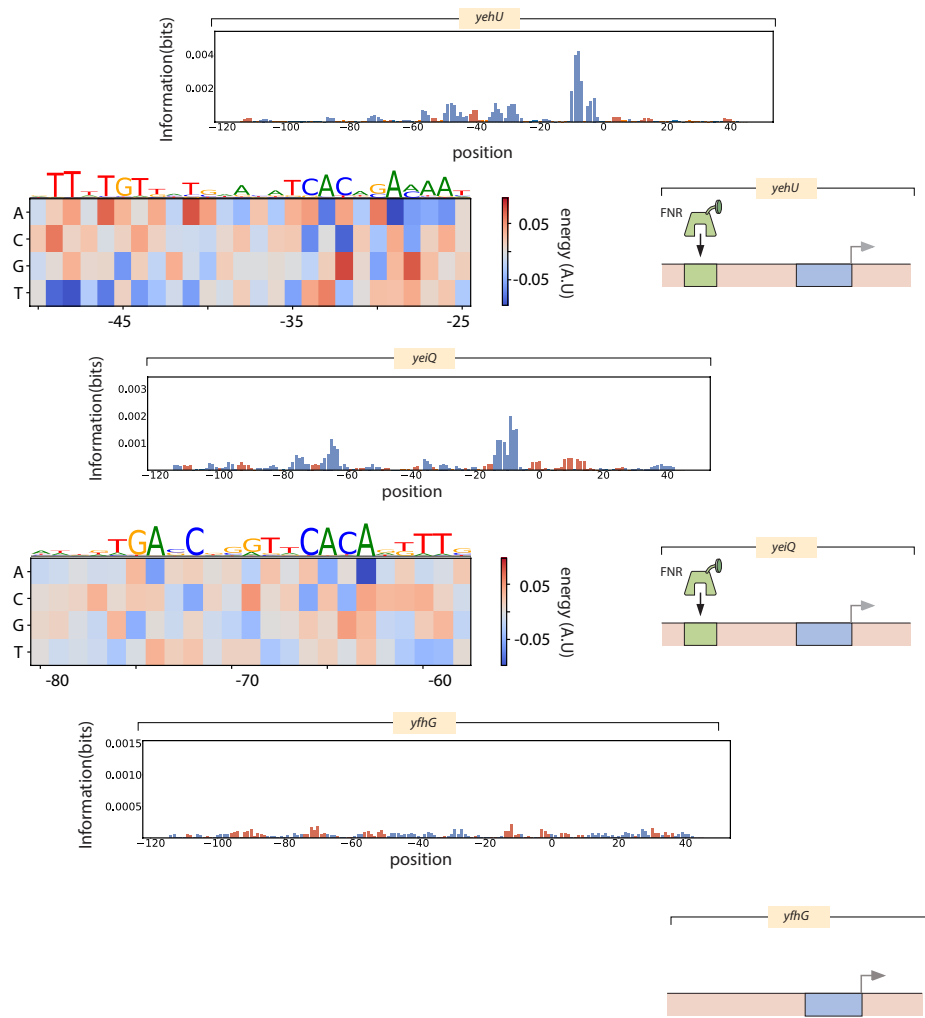


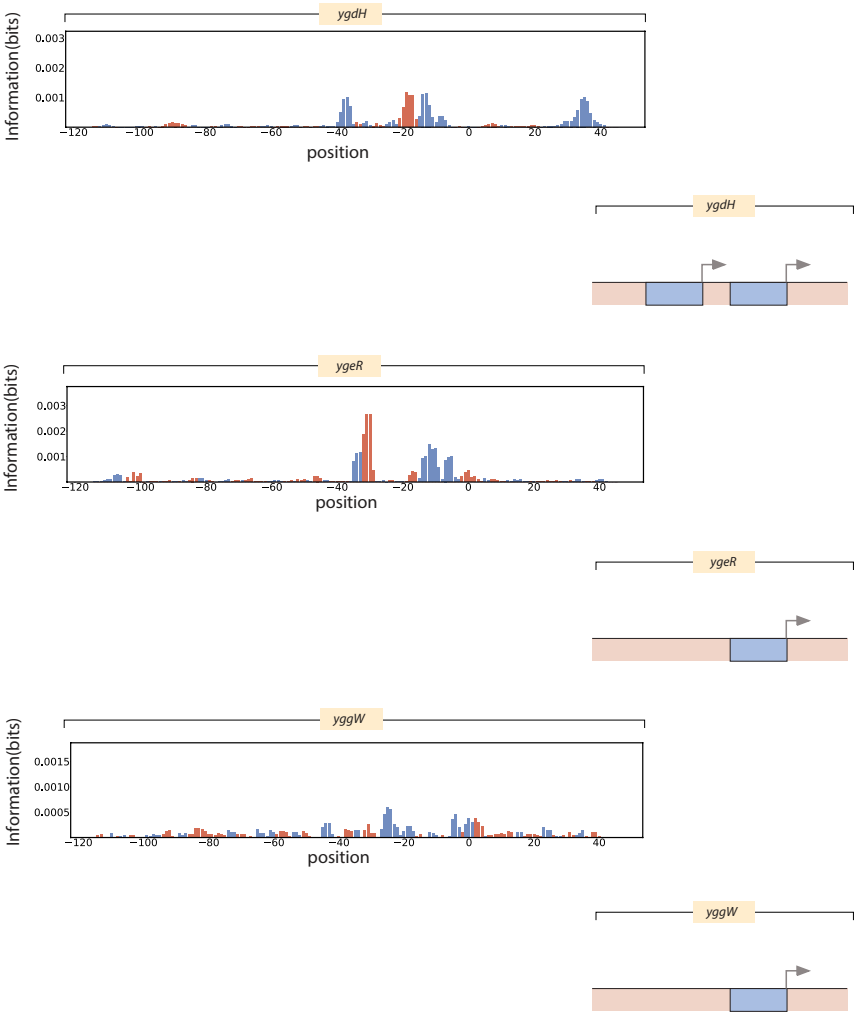


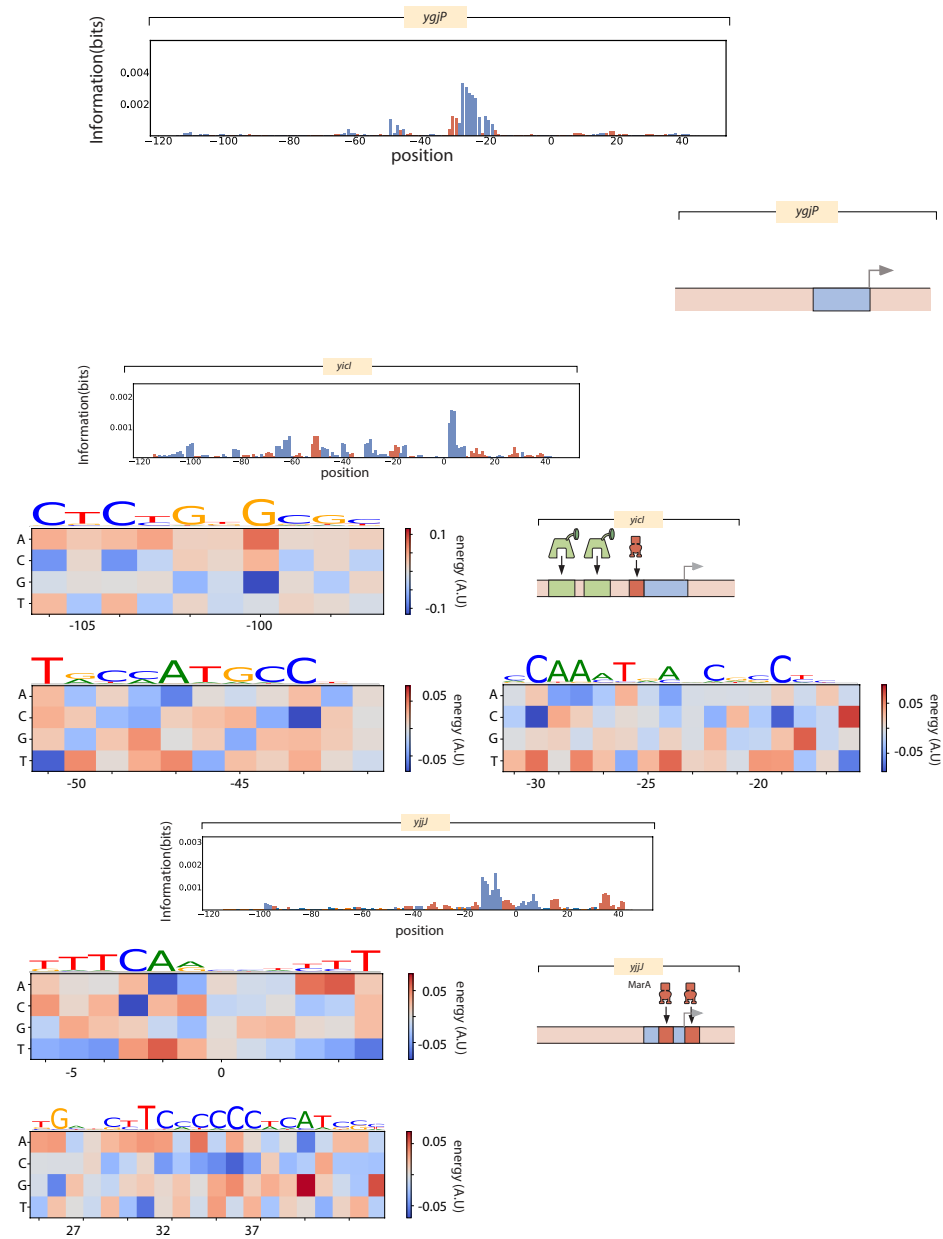


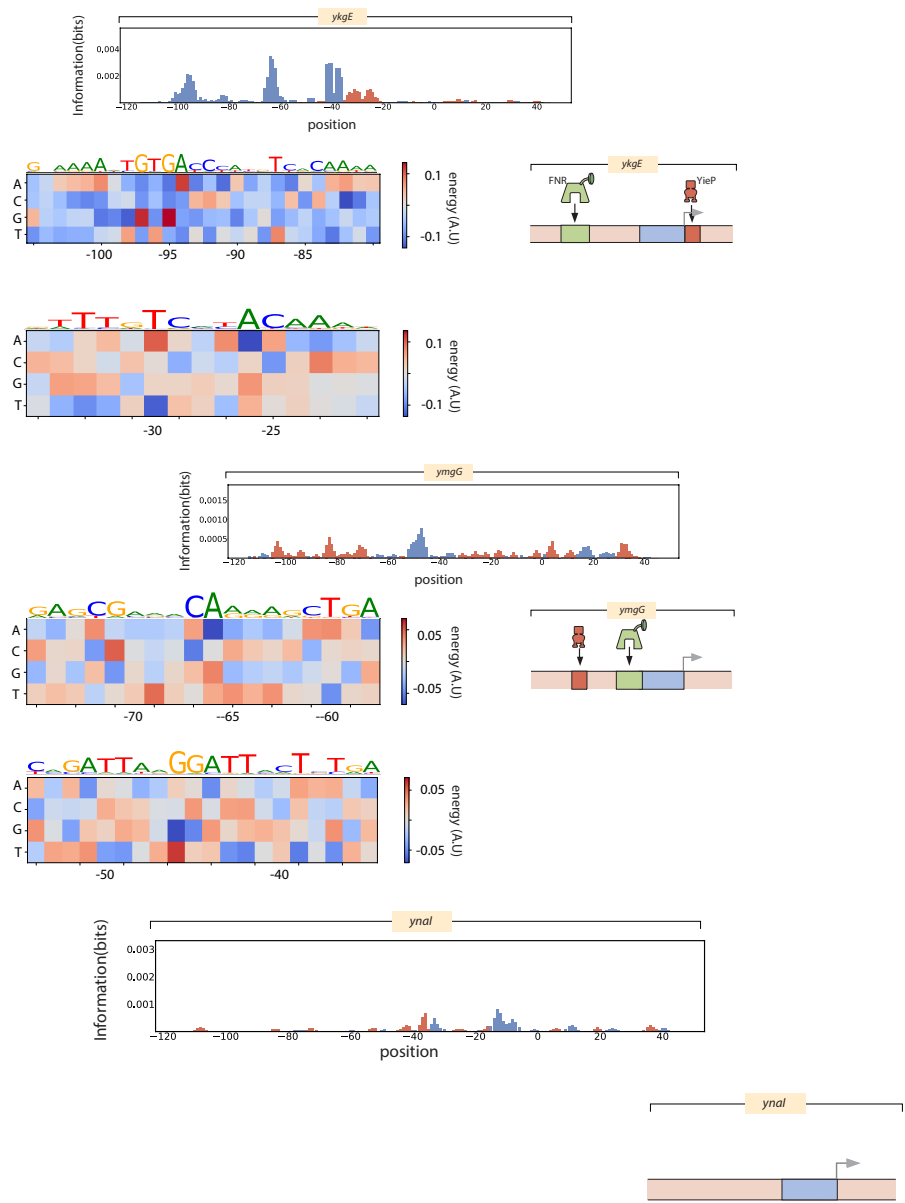


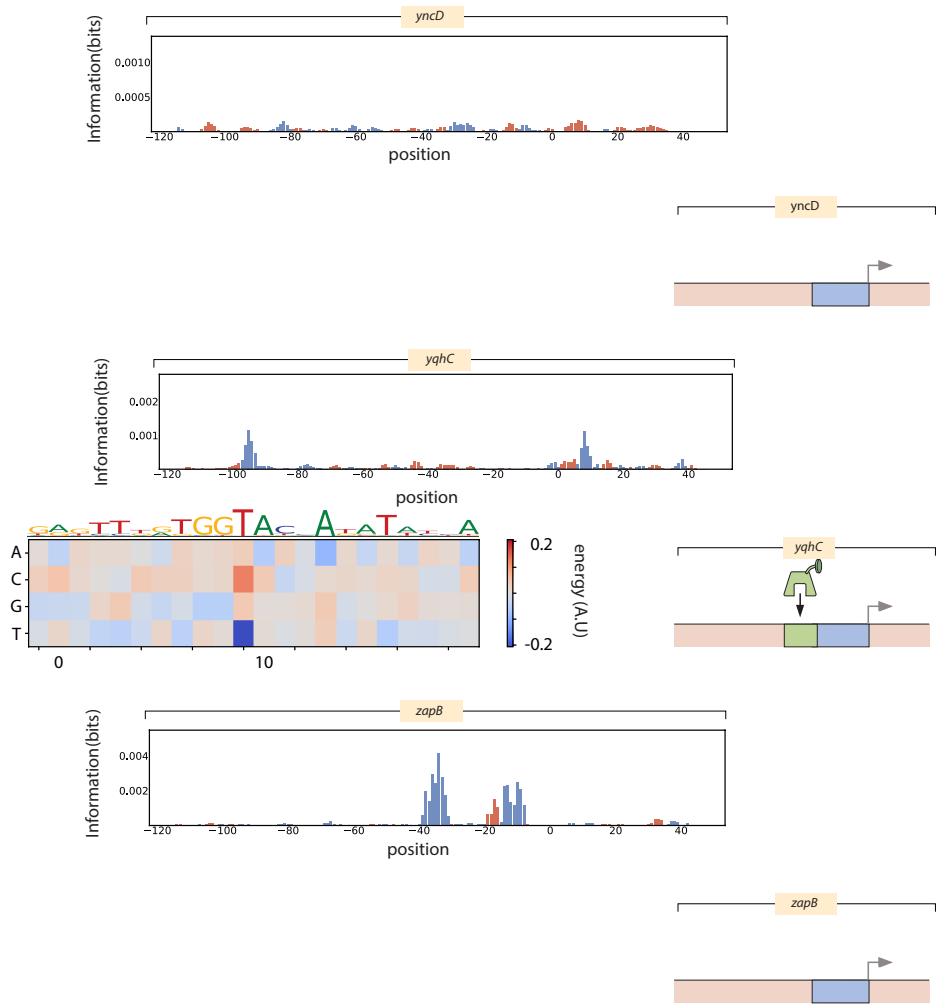












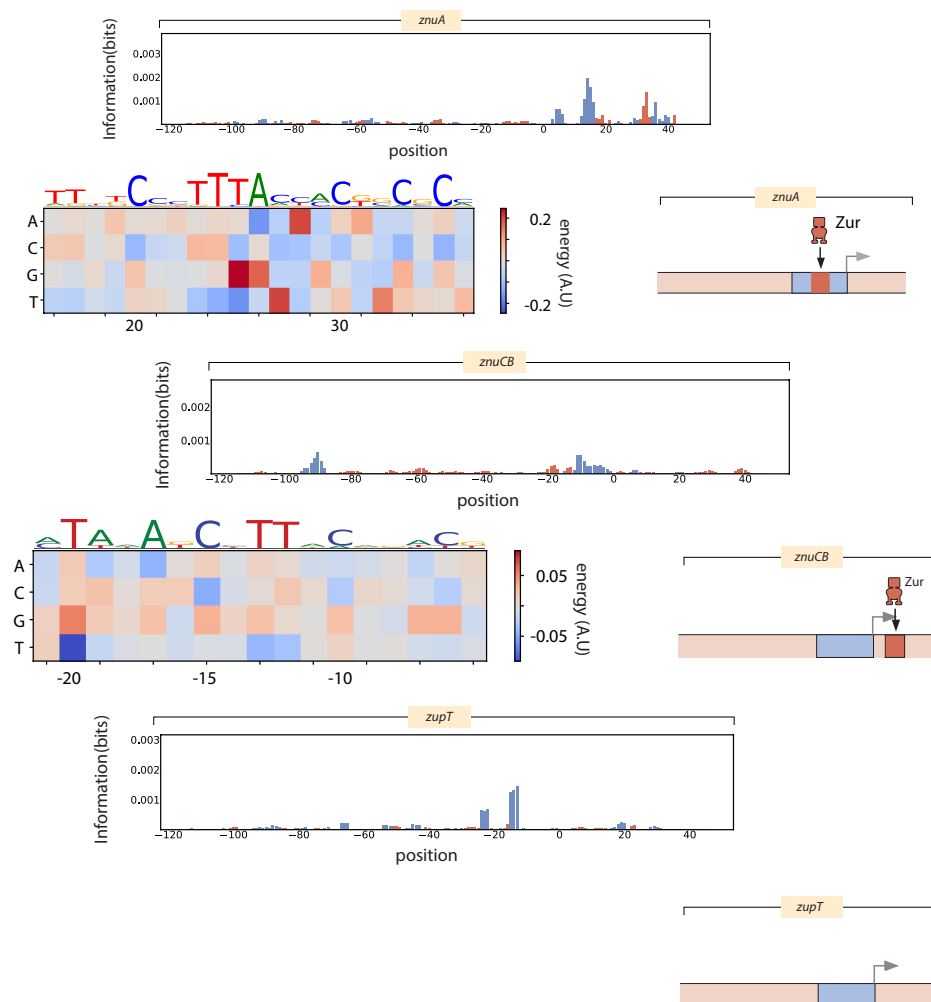


Figure B.13: Data from Reg-Seq

Information footprints (top panels), regulatory cartoons (bottom right panel), and energy matrices (left panels).

BIBLIOGRAPHY

- Al Mamun, Abu Amar M. et al. (Dec. 2012). “Identity and function of a large gene network underlying mutagenic repair of DNA breaks”. eng. In: *Science (New York, N.Y.)* 338.6112, pp. 1344–1348. doi: 10.1126/science.1226683.
- Barnes, Stephanie L. et al. (2019). “Mapping DNA sequence to transcription factor binding energy *in vivo*”. In: *PLoS Computational Biology* 15.2, pp. 1–29. doi: 10.1371/journal.pcbi.1006226.
- Belliveau, Nathan M. et al. (2018). “Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.21, E4796–E4805. doi: 10.1073/pnas.1722055115.
- Compan, Inès and Danlèle Touati (1994). “Anaerobic activation of *arcA* transcription in *Escherichia coli*: roles of Fnr and ArcA”. en. In: *Molecular Microbiology* 11.5, pp. 955–964. doi: 10.1111/j.1365-2958.1994.tb00374.x.
- Crooks, Gavin E. et al. (June 2004). “WebLogo: A Sequence Logo Generator”. en. In: *Genome Research* 14.6, pp. 1188–1190. doi: 10.1101/gr.849004.
- Easton, A M and S R Kushner (Dec. 1983). “Transcription of the *uvrD* gene of *Escherichia coli* is controlled by the *lexA* repressor and by attenuation.” In: *Nucleic Acids Research* 11.24, pp. 8625–8640. doi: 10.1093/nar/11.24.8625.
- Forcier, Talitha L et al. (2018). “Measuring cis-regulatory energetics in living cells using allelic manifolds”. en. In: p. 28.
- Garcia and Phillips (July 2011). “Quantitative dissection of the simple repression input-output function”. en. In: *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–12178. doi: 10.1073/pnas.1015616108.
- Gupta, Shobhit et al. (2007). “Quantifying similarity between motifs”. In: *Genome Biology* 8.2. doi: 10.1186/gb-2007-8-2-r24.
- Ireland, William T. and Kinney (May 2016). “MPAthic: Quantitative Modeling of Sequence-Function Relationships for massively parallel assays”. en. In: doi: 10.1101/054676.
- Jain, Chaitanya (Feb. 2008). “The *E. coli* RhIE RNA helicase regulates the function of related RNA helicases during ribosome assembly”. eng. In: *RNA (New York, N.Y.)* 14.2, pp. 381–389. doi: 10.1261/rna.800308.
- Keseler, Ingrid M. et al. (Jan. 2013). “EcoCyc: fusing model organism databases with systems biology”. en. In: *Nucleic Acids Research* 41.D1, pp. D605–D612. doi: 10.1093/nar/gks1027.
- Kinney (2008). “Biophysical models of transcriptional regulation from sequence data”. en. In: p. 124.

- Kinney and Atwal (Dec. 2013). “Parametric inference in the large data limit using maximally informative models”. en. In: *arXiv:1212.3647 [math, q-bio, stat]*. arXiv: 1212.3647.
- Kinney, Anand Murugan, et al. (2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.20, pp. 9158–9163. DOI: 10.1073/pnas.1004290107.
- Kumar, Rahul and Kazuyuki Shimizu (2011). “Transcriptional regulation of main metabolic pathways of *cyoA*, *cydB*, *fnr*, and *fur* gene knockout *Escherichia coli* in C-limited and N-limited aerobic continuous cultures”. en. In: *Microbial Cell Factories* 10.1, p. 3. DOI: 10.1186/1475-2859-10-3.
- Magoc and Salzberg (Nov. 2011). “FLASH: fast length adjustment of short reads to improve genome assemblies”. en. In: *Bioinformatics* 27.21, pp. 2957–2963. DOI: 10.1093/bioinformatics/btr507.
- Mendoza-Vargas, Alfredo et al. (Oct. 2009). “Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*”. en. In: *PLoS ONE* 4.10. Ed. by Chad Creighton, e7526. DOI: 10.1371/journal.pone.0007526.
- Monsieurs, Pieter et al. (Apr. 2005). “Comparison of the PhoPQ regulon in *Escherichia coli* and *Salmonella typhimurium*”. eng. In: *Journal of Molecular Evolution* 60.4, pp. 462–474. DOI: 10.1007/s00239-004-0212-7.
- Murray, Sean et al. (Oct. 2001). “Extragenic Suppressors of Growth Defects in *msbB* *Salmonella*”. en. In: *Journal of Bacteriology* 183.19, pp. 5554–5561. DOI: 10.1128/JB.183.19.5554-5561.2001.
- Neal, Radford M (1993). “Probabilistic Inference Using Markov Chain Monte Carlo Methods”. en. In: p. 144.
- Partridge, Jonathan D. et al. (2009). “NsrR targets in the *Escherichia coli* genome: new insights into DNA sequence requirements for binding and a role for NsrR in the regulation of motility”. en. In: *Molecular Microbiology* 73.4, pp. 680–694. DOI: 10.1111/j.1365-2958.2009.06799.x.
- Price, Morgan N. et al. (May 2018). “Mutant phenotypes for thousands of bacterial genes of unknown function”. en. In: *Nature* 557.7706, pp. 503–509. DOI: 10.1038/s41586-018-0124-0.
- Raetz, Christian R.H. et al. (June 2007). “Lipid A Modification Systems in Gram-Negative Bacteria”. In: *Annual Review of Biochemistry* 76.1, pp. 295–329. DOI: 10.1146/annurev.biochem.76.010307.145803.
- Rhee, Kyu Y., Donald F. Senear, and G. Wesley Hatfield (May 1998). “Activation of Gene Expression by a Ligand-induced Conformational Change of a Protein-DNA Complex”. en. In: *Journal of Biological Chemistry* 273.18, pp. 11257–11266. DOI: 10.1074/jbc.273.18.11257.

- Santos-Zavaleta, Alberto et al. (2019). “RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli K-12*”. In: *Nucleic Acids Research* 47, pp. 212–220. DOI: 10.1093/nar/gky1077.
- Schmidt, Alexander, Karl Kochanowski, Silke Vedelaar, Erik Ahrné, et al. (2015). “The quantitative and condition-dependent *Escherichia coli* proteome”. In: *Nature Biotechnology* 34.1, pp. 104–110. DOI: 10.1038/nbt.3418.
- Schmidt, Alexander, Karl Kochanowski, Silke Vedelaar, Erik Ahrné, et al. (Jan. 2016). “The quantitative and condition-dependent *Escherichia coli* proteome”. en. In: *Nature Biotechnology* 34.1, pp. 104–110. DOI: 10.1038/nbt.3418.
- Schneider, Thomas D. et al. (Apr. 1986). “Information content of binding sites on nucleotide sequences”. en. In: *Journal of Molecular Biology* 188.3, pp. 415–431. DOI: 10.1016/0022-2836(86)90165-8.
- Stormo, G. D. (Jan. 2000). “DNA binding sites: representation and discovery”. en. In: *Bioinformatics* 16.1, pp. 16–23. DOI: 10.1093/bioinformatics/16.1.16.
- Suzuki, Masashi (2003). “The DNA-binding specificity of eubacterial and archaeal FFRPs”. In: *Proceedings of the Japan Academy, Series B* 79B.7, pp. 213–222. DOI: 10.2183/pjab.79B.213.
- Tareen, Ammar and Kinney (Dec. 2019). “Biophysical models of cis-regulation as interpretable neural networks”. en. In: *bioRxiv*, p. 835942. DOI: 10.1101/835942.
- Valens, Michèle, Axel Thiel, and Frédéric Boccard (Sept. 2016). “The MaoP/maoS Site-Specific System Organizes the Ori Region of the *E. coli* Chromosome into a Macrodomain”. In: *PLoS Genetics* 12.9. DOI: 10.1371/journal.pgen.1006309.
- Zwir, Igor et al. (May 2012). *The promoter architectural landscape of the Salmonella PhoP regulon*. en. DOI: 10.1111/j.1365-2958.2012.08036.x.

INDEX

B

bibliography

by chapter, 19, 45, 91, 126, 181

F

figures, 3, 7–11, 16, 17, 19, 26, 28, 31, 32, 35, 37, 40, 53, 56, 58, 60, 61, 63, 65,
66, 68, 70, 72, 79, 90, 105, 106, 112, 114, 115, 131–134, 142, 147, 149,
156, 158, 160, 161, 163, 167, 210

T

tables, 34, 46, 113, 130, 145, 163, 166